INDEXING POPULATIONS

BACKGROUND

Characterizing populations is a general problem encountered in many areas of science and technology. For example, determining the types molecules present in a sample and their amounts is a primary goal of much of analytical chemistry and biochemistry. Clinical diagnostic chemistry represents one particular class of such analytical applications. Many of the important analytical techniques aim to detect or determine the amount of only one specific constituent in a mixture and they generally rely on purifying the component of interest away from other members of the population, to a greater or lesser extent. Tests that measure cholesterol in blood provide a good example in this regard. In addition, such clinically prevalent one component analytical tests typically are developed with a thorough understanding and a supply of purified and well characterized preparations of the target.

Tests that determine only one or even a few, components of a mixture represent a small step toward a broader goal of detecting and quantifying many or even all of the components of a population, even when many of them have not been identified, have not been characterized and are not available for study. The difficulty and the promise of methods that can profile little-characterized populations is exemplified by the recent development of methods to determine virtually all of the different types of mRNAs in a sample population and the amount of each type. Recently developed single nucleotide polymorphism ("SNP") screening methods and their application to pharmacogenomic studies provide a similar illustration of the advantages of techniques that make it practical to profile large, complex, poorly characterized populations. Similar profiling methods for a variety of other populations and purposes also hold great promise.

In general, such methods must solve two problems: how to distinguish different sub-populations of interest in a population and how to count the number of individuals in each sub-population. For instance, general methods for characterizing or profiling a population of individuals in this way must distinguish different sub-populations in the population, generally defined by distinguishably different types, and count the number of individuals in each sub-population. The resulting list of sub-population and or distinguishable types of individuals and the number of individuals of each in the population characterizes and provides a profile of a the population that can be used to compare it to other populations. Generating substantially complete profiles of this type for large populations containing many sub-populations of interest is an unmet challenge, particularly when few of the sub-populations have been characterized.

The general problem of profiling large complex populations without prior characterization of its sub-populations, and the limitations of current technology, is illustrated by- several recently developed techniques for genome-wide gene-expression profiling. All of these techniques are designed to determine all-at-once the expression of substantially all the genes in cells in a sample. A typical cell contains a population of mRNAs that reflects the activity of its genes. The mRNA population is made up of different types of mRNAs present in different numbers. Different types of mRNA in the population have sequences that differ from one another whereas, by definition, copies of the same type of mRNA have the same sequence as one another. A population of mRNAs can be characterized, profiled and even defined, by a list of different types of mRNA and how many copies of each type it contains. In principle, the expression of all genes in a cell can be determined by identifying and quantifying the different mRNAs it contains. Such mRNA profiling can be useful to characterize and understand differences in gene expression between, for instance, pathogenic cells and their normal counterparts, such as cancer cells, cells infected by a virus, cells affect by other dysfunctions and their normal counterparts. Profiling mRNA populations and gene expression thus can provide insights for understanding normal and dysfunctional cell function and can aid the development of methods and treatments to prevent, cure or ameliorate disease.

Several of the techniques that have been developed thus far to profile mRNA populations without prior characterization have several features in common. Foremost, they all separate the different types of mRNA in the sample from one another so that they can be detected and quantified independently. In addition, the techniques generally first transform the mRNA into cDNA. They then use tags (also referred to herein as indexing sequences, among other things) to recognize and separate different types of cDNA. The tags typically are short sequences, such as dinucleotide anchors on first strand reverse transcriptase primers, random primers, arbitrary sequence primers, restriction enzyme cleavage sites and overhangs produced by cleavage by Type IIS restriction enzymes. The tags are used to separate different types of cDNAs into sub-populations. Techniques that use Type IIS enzymes in this way are described by, for instance, Smith *et al.*, *PCR Methods and Applns* 2: 21-27 (1992) and Unrau *et al.*, *Gene* 145: 163-169 (1994), each of which is incorporated herein by reference in its entirety.

Tagging and separation often are reiterated to yield relatively small sub-populations of 50 to 100 different cDNA species that then are separated from one another by slab or capillary electrophoresis. Each resulting gel band or CE peak represents at least one of the cDNA species in the sub-populations defined by the tagging procedure; although, similarly sized but different types of cDNAs that occur in the same sub-population will overlap and may be seen as a single species rather than two or three or more species of the same size. The amount of material in each band or peak provides a measurement of the amount of the corresponding mRNA sub-population in the sample. All the bands or peaks in all the lanes or CE runs taken together can represent a substantial fraction of all the different mRNAs in the original sample.

READS, for instance, first subdivides the mRNA from a cell into sub-populations using different primers to synthesize different sub-populations of cDNA. See, for instance, U.S. patent. No. 5,712,126 of Weissman *et al.* on *Analysis of Gene Expression By Display of 3'-end Restriction Fragments of cDNA* which is incorporated by reference herein in its entirety. The cDNA of each sub-population then is further sub-divided by digestion with different restriction enzymes. 3' end fragments are specifically amplified from the digests and resolved by size in sequencing gels. Typically, the first subdivision is carried out using 12 different anchor primers for first strand synthesis. Each of the resulting 12 sub-populations of cDNA is then typically sub-divided into 30 aliquots and digested with 30 different restriction enzymes. The 3' end-fragments from each of the 240 sub-populations are specifically amplified and then separated by size on sequencing gels. Assuming 50 fragments are resolved in each lane on average, READS carried out with 12 primers and 30 enzymes provides about 16,500 fragments that represent the mRNA population. Such a display can be produced using about 15 gels. The different restriction primers and enzymes sample the mRNA population independently of one another. The displays, therefore, statistically represent the mRNA population. Some mRNAs will not be represented by any fragments in the gels. Others will be represented by one fragment. Still others will be represented by two or more fragments. The larger the number of different primers and different enzyme recognition sites the greater the statistical accuracy of the method.

Another technique for indexing that utilizes restriction enzyme cleavage followed by electrophoresis to resolve individual fragment species is described in U. S. patent No. 5,8-14,445 to Balyavsky *et al.*, which is incorporated herein by reference in its entirety.

The "differential display" technique of Liang and Pardee, employs random primers to amplify in a complex mRNA or cDNA population a small number of fragments that can be resolved in a single gel lane. The method is described in Liang *et al., Science 257*: 967-971 (1992), which is incorporated herein by reference in its entirety. Related methods, based on non-specifically primed amplification followed by gel electrophoresis to resolve individual amplification products apparently are used by the companies Keygene and Curagen as described in, for instance, International Publication Number WO 93/06239 of Zabeau *et al.* (Keygene) and U.S. patent number 5,871,697 of Rothberg *et* al. (Curagen) which both are incorporated herein by reference in their entireties.

While most profiling methods rely on size separation techniques such as slab and capillary electrophoresis, as described above, to resolve and quantify fragments that represent sub-populations of interest, a few profiling methods involve other techniques.

One method, called SAGE, uses a Type IIS restriction enzyme to make short 3' end "tag" fragments from cDNAs in a population. The tags ligated together into dimers, the dimers are amplified and the amplicons are concatenated and cloned. After re-isolation from clones, the catenates are sequenced. Gene expression is profiled by identifying the different tag sequences in the catenates, which can be deduced directed from the dimer structures, and by counting the number of times each sequence occurs. Some 60,000-90,000 tags must be sequences for a statistically accurate portraits of gene expression in a human cell. Using a system that can sequence 1,000 tags per gel (about 20,000 base pairs), SAGE requires about 60 to 90 gels per expression profile. SAGE is described in, for instance, U.S. patent numbers 5,695,937 and 5,866,330, both for a *Method for Serial Analysis of Gene Expression* and both to Kinkier *et al.*, each of which is incorporated herein by reference in its entirety.

Another approach to expression profiling uses positional arrays of probes on a substrate. All the probes on the substrate are contacted with a sample and processed at once. Results are determined by reading the signal at each probe position. The method has produced striking results in profiling gene expression in yeast using both long cloned DNA and short synthetic oligonucleotides as probes. Using short oligonucleotides as probes required about 250,000 different probes to accurately quantify expression of 6,200 putative genes in yeast. Further development is necessary before the method can be applied to assay expression of 15,000 genes typically active in a human cell. Furthermore, arrays are expensive to synthesize and they can be used only in conjunction with expensive, complex, specialized and dedicated equipment for processing and for reading out results. Approaches based on solid substrate arrays in this regard are reviewed in Lipshulz *et al., Nature Genetics 2-1:* 20-24 (1999), for instance.

Although the foregoing discussion is directed to the illustrative example of profiling mRNA populations, the same problems are posed by current techniques being developed to profile other types of populations. Indeed, effective profiling techniques would be useful for populations of many other types. For instance such techniques would be useful for profiling populations of genomic fragments for ordering, for profiling cDNA or genomic fragments to detect allelic variations or to characterize a genome against a panel of SNP probes. (Regarding SNP analysis see, for instance, Vos *et al., Nuc. Acids Res.* 21: 4407-4414 (1995) which is incorporated herein by reference in its entirety.) Such techniques also could be used to characterize populations chemical synthesis, such as compounds after screening a combinatorial or other compound library. Such techniques in principle could be used to characterize, profile or define populations of virtually any molecule, macromolecular assembly or complex, mixtures of cells, or virtually any other substance that has an identifiable tag.

All of these techniques and methods involve manually intensive, time consuming and difficult procedures that cannot be completely automated. Methods that rely on size separation for instance incur significant disadvantages of gel or column preparation, specialized instrumentation for running and reading gels or columns, long resolution times of electrophoretic separations, manual steps of gel and column preparation and the lack of fully automated techniques for electrophoretic separations. Arrays suffer from significant drawbacks as well. They are relatively inflexible, expensive, can be made only with highly sophisticated, very expensive equipment and skills, can be used only with specially designed, expensive equipment and instrumentation, and currently suffer serious limitations in reliability, precision and accuracy. Thus, unfortunately, current methods for profiling are limited, cumbersome, time-consuming, expensive, require difficult, expensive, time consuming and imprecise size separation techniques, such as slab gel electrophoresis, capillary electrophoresis, HPLC and the like, cannot be carried out in homogeneous media, and 'cannot be fully automated, among other problems, drawbacks and limitations. Hence, it has not been possible to fully deploy indexing methods for studying large populations, how they change or how they differ in different sources.

Therefore there exists a need for, among other things, methods, reagents, devices, apparatuses and the like for profiling populations, such as populations of molecules, particularly combinatorial polymers, such as mRNA and DNA, for, in particular, expression profiling and genetic screening, among others, that are simple, relatively fast and easy to automate.

SUMMARY

It is among the objects of the invention, therefore, to provide, among others, methods, processes, devices, machines, apparatus, manufactures and compositions of matter for, among others, analyzing, detecting, determining, characterizing, counting and/or quantifying constituents, individuals, items, members, objects and the like of given species, type, groups, sub-populations and the like in a population, and to, characterize, profile and/or define populations thereby.

It is a particular object of the invention in this regard to provide, among other things, methods, processes, devices, machines, apparatus, manufactures and compositions of matter for analyzing, characterizing, profiling, determining and/or defining populations by determining different species in the populations and the number of individuals of each species. In this regard, a more particular object of the invention relates to populations of a plurality of discrete individuals, wherein each individual can be classified as one or more of a plurality of types, and to characterizing such discrete populations by determining the number of individuals of each type. In a yet more particular aspect, the invention relates to populations of molecules and to characterizing molecule populations by detecting the presence and determining the amount of different molecular species in the population. A yet more highly particular aspect of the invention in this regard relates to profiling polynucleotide populations by detecting the presence and determining the amount of different polynucleotide species in the population. Certain very highly particular aspects of the invention in this regard relate to profiling populations of genomic or DNA for genetic screening, such as SNP screening, and to gene expression profiling of mRNA and cDNA population.

It is another particular aspect of the invention to provide, among other things, methods, processes, devices, machines, apparatus, manufactures and compositions of matter for analyzing, characterizing, profiling, determining and/or defining populations by determining many of the different species in the population and the number of individuals of each of the species so determined. In a particular aspect in this regard the invention relates to among other things, methods, processes, devices, machines, apparatus, manufactures and compositions of matter for analyzing, characterizing, profiling, determining and/or defining populations by determining a statistically determined number of species and the number of individuals of each species, wherein the number of species determined is statistically determined to be representative to a given degree of accuracy of a chosen fraction of all the species that can occur in the population. Particular aspects of the invention in this regard relate to the same in which substantially all the different species that can occur in the population are determined.

Further particular illustrative aspects and objects of the invention are set out in the following numbered paragraphs.

1. A method for indexing a population, comprising determining products of the reaction of a population to be indexed with each of a plurality of different indexing reagents, wherein

each reaction product is indicative of a different cognate sub-population of individuals,

the population is indexed by the constituency of sub-populations thus determined;

the reactions are conducted in homogeneous media, and

the reaction products are determined in homogeneous media.

2. A method for indexing a population, comprising determining products of the reaction of a population to be indexed with each of a plurality of different indexing

reagents, wherein

each reaction product is indicative of a different cognate sub-population of individuals,

the population is indexed by the constituency of sub-populations thus determined;

the reaction products are determined without physically separating determined products from the reactions.

3. A method for indexing a population, comprising determining products of the reaction of a population to be indexed with each of a plurality of different indexing reagents, wherein

each reaction product is indicative of a different cognate sub-population of individuals,

the population is indexed by the constituency of sub-populations thus determined;

the reactions are conducted in homogeneous media, and

the reaction products are determined without physically separating determined products from the reactions.

4. A method for indexing a population, comprising determining products of the reaction of a population to be indexed with each of a plurality of different indexing reagents, wherein

each reaction product is indicative of a different cognate sub-population of individuals,

the population is indexed by the constituency of sub-populations thus determined;

the reactions are conducted in homogeneous media;

the reaction products are determined in homogeneous media, and

the reaction products are determined without physically separating the determined products from the reactions.

5. A method according to any of the foregoing numbered paragraphs, wherein the reactions are determined over time and the resulting kinetic data is used to determine the reaction products.

6. A method according to paragraph 5, wherein the reactions are monitored continuously.

7. A method according to paragraph 5, wherein the reactions are monitored discontinuously.

8. A method according to any of the foregoing numbered paragraphs, wherein reactions of two or more indexing reagents with the population are multiplexed together.

9. A method according to any of the foregoing numbered paragraphs, wherein a substantial fraction of the possible sub-populations of a given type are thus determined.

10. A method according to paragraph 9, wherein the substantial fraction is any of or range between any two of 25%, 30%, 35%, 40%, 45% 50%, 55%, 60%, 65%, 70%, 75%, 80%, 82%, 84%, 85%, 86%, 88%, 90%, 92%, 94%, 95%, 96%, 97%, 98%, 99% or more than 99%.

11. A method according to paragraph 9 or 10, wherein substantially all of the sub-population are determined.

12. A method according to any of the foregoing numbered paragraphs, wherein the population is a population of polynucleotides and the sub-populations are sequence-specific sub-populations.

13. A method according to paragraph 12, wherein sub-populations represents the expression of genes.

14. A method according to -paragraph 12 or 13, wherein the population is a population of mRNAs.

15. A method according to -paragraph 12 or 13, wherein the population is a

representative of mRNAs.

16. A method according to -paragraph 12 or 13, wherein the population is a population of cDNAs.

17. A method according to any of the foregoing numbered paragraphs, wherein the indexing reagents comprise sequence-specific indexing probes.

18. A method according to paragraph 17, wherein the sequence specific probe is specific for a sequence n Oases long, where n is 1, 2, 3, 4, 5, 6, 7 or 8.

19. A method according to paragraph 18, wherein the sequence specific probe is specific for a sequence n bases long, where n is 1, 2, 3, 4, 5 or 6.

20. A method according to paragraph 19, wherein the sequence specific probe is specific for a sequence n bases long, where n is 2, 3, 4 or 5.

21. A method according to any of paragraphs 17 through 20, wherein the sequence-specific probe is h bases long, each different sequence possible for a sequence n bases long is specific for a sub-population to be determined, and indexing reactions between the population and indexing reagents that comprise a substantial fraction of all the possible sequences n bases long a re determined.

22. A method according to paragraph 9, wherein the substantial fraction is any of or range between any two of 25%, 30%, 35%, 40%, 45% 50%,55%, . 60%, 65%, 70%, 75%, 80%, 82%, 84%, 85%, 86%., 88%, 90%, 92%, 94%, 95%, 96%, 97%, 98%, 99% or more than 99%.

23. A **method according to paragraph 21,** wherein substantially all of the sub-populations defined by the probe sequences are determined.

24. A method according to paragraph 21, wherein all of the sub-populations defined by the probe sequences are determined.

25. A method according to any of the foregoing numbered paragraphs, wherein

the indexing reagents comprise amplification primers.

26. A method according to any of the foregoing numbered paragraphs, wherein the probes are stand-d is placement indexing adaptors.

27. A method according to paragraph 25, comprising:

(A) distributing strand-displacement indexing adaptors into discrete containers;

(B) contacting each of the strand-displacement indexing adaptors in the containers with an aliquot of a sample population of polynucleotides;

(C) carrying out strand-displacement indexing reactions between the sample polynucleotides and the indexing adaptors in each container;

(D) amplifying by PCR the polynucleotides that form strand displaced structures with indexing adaptors in each container;

(E) quantifying the amount of amplified polynucleotide in each container, thereby determining the absence or the amount of cognate polynucleotides in the population, and

(F) indexing the population by the absence, the presence, and amount of sample polynucleotides cognate to the indexing adaptors in each of the containers.

28. A method according to paragraph 25, wherein the indexing sequence comprises a Type 11 restriction cite contiguous with a five nucleotide base-pair sequence.

29. A device or set of devices for indexing a population, comprising a plurality indexing reagents disposed in a plurality of containers, wherein the indexing reagent in each container is specific for a cognate sub-population and the plurality of indexing reagents in the device comprises reagents specific for a substantial fraction of all the sub-populations of a given type that can be indexed in the population.

30. A device according to paragraph 29, wherein the substantial fraction is any of or range between any two of 25%, 30%, 35%, 40%, 45% 50%, 55%, 60%, 65%, 70%, 75%, 80%, 82%, 84%, 85%, 86%, 88%, 90%, 92%, 94%, 95%, 96%, 97%, 98%, 99% or more than 99%.

31. A device according to paragraph 29, wherein substantially all of the sub-populations of the given type are determined.

32. A device according to paragraph 29, wherein the containers are wells of one or more micro-titer plates.

33. A device according to any of paragraphs 29-32, wherein the indexing reagents comprise sequence-specific indexing probes.

34. A device according to paragraph 33, wherein the sequence specific probe is specific for a sequence n bases long, where n is 1, 2, 3, 4, 5, 6, 7 or 8.

35. A device according to paragraph 32, wherein the sequence specific probe is specific for a sequence n bases long, where n is 1, 2, 3, 4, 5 or 6.

36. A device according to paragraph 33, wherein the sequence specific probe is specific for a sequence n bases long, where n is 2, 3, 4 or 5.

37. A device according to any of paragraphs 29 through 33, wherein the sequence-specific probe is n bases long, each different sequence possible for a sequence n bases long is specific for a sub-population to be determined, and the plurality of indexing reagents comprise indexing reagents with a substantial fraction of all the possible sequences n bases long.

38. A device according to paragraph 37, wherein the substantial fraction is any of or range between any two of 25%, 30%, 35%, 40%, 45% 50%, 55%, 60%, 65%, 70%, 75%, 80%, 82%, 84%, 85%, 86%, 88%, 90%, 92%, 94%, 95%, 96%, 97%, 98%, 99% or more than 99%, substantially all and all.

DESCRIPTION OF TERMS

The following description is provided to illuminate some terms and phrases used herein to describe aspects of the invention. The brief explanations below are meant to be informative; but, they are not intended to be exhaustive. The meaning of the terms here can be fully understood only by reading the entire disclosure with knowledge of those skilled in the arts to which the invention pertains. The illustrative explanation set out immediately below as an aid to understanding the disclosure thus should not be taken as limitations on the meaning of terms that unduly limit the scopes of the invention herein disclosed.

ABSENCE The term absence generally is used herein to refer either to complete absence or to presence in an amount less than an amount required for detection by one or more detection methods.

AMPLIFICATION The term amplification generally is used herein to refer to processes for increasing the number of copies of the object of amplification. For instance, certain preferred embodiments relate to the amplification of polynucleotides using PCR, which can increase the number of copies of the amplified polynucleotides by several orders of magnitude. In certain preferred embodiments of the invention is this regard, amplification is carried out so that only indexed individuals are amplified and only amplified products are detected. Amplification may be linear, exponential or a combination of the two. Linear amplification may be accomplished using a promoter system, such as bacteriophage-derived promoter systems, such as T3 and T7 promoter systems, and by linear PCR, for instance. Exponential amplification may be accomplished using PCR or by a variety of other methods.

AMPLIFICATION PRIMER The term amplification primer s used herein generally to refer to an oligonucleotide designed to hybridize to a specific sequence in a template and to provide a T-OH that can be extended opposite the template strand by a template-dependent polymerase, such as a reverse transcriptase or a DNA-dependent DNA polymerase. Primers, in the form added to a reaction, can be a single molecular specie of a single defined sequence or a mixture of any number of species of molecules that differ in sequence. Primers may be made entirely of naturally occurring constituents, entirely of non-naturally occurring constituents, such as PNAs, or they may be made partially of naturally occurring and partially of non-naturally occurring constituents. Amplification primers may comprise sequences and moieties in addition to the sequence or sequences specially involved in hybridizing to the template. Among the additional sequence and moieties are detectable labels, binding moieties, indexing sequences and/or sequences or moieties for binding or attaching the same.

AMPLIFICATION SEQUENCE The phrase amplification sequence is used herein generally to refer to a sequence that can be used to amplify a polynucleotide. Often the sequence is comprised in an adaptor that is attached to the end of the amplification target, such as an indexing adaptor that hybridizes and is ligated to a target polynucleotide. The amplification sequence for a given amplification primer may be complementary to that of the primer. However, it may also be the same as that of the primer. This will be case, for instance, when a partially double-stranded adaptor is ligated to the end of a target polynucleotide so that, initially, the only polymerization step that can occur proceeds from a 3' end on one strand of the target polynucleotide using the opposite strand of the adaptor as the template. This initial step produces a primer-complementary sequence in the extended strand of the target.

ANCHOR The term anchor as used herein generally refers to a sequence of one or more nucleotides that defines the position of hybridization of a primer or an adaptor to a target polynucleotide. Anchors are useful for parsing populations. They also are used for other reasons because they provide better specificity and or more efficient priming than other primers that might be available for a given application. Perhaps the most common example of an anchor is the dinucleotide anchor at the 3' end of the oligo dT primer frequently used in first strand cDNA synthesis. The dinucleotide anchor at the 3' end of the oligo dT track forces the primers to hybridize to the Same spot in all mRNAs: at the junction of the 3' end of the mRNA with the polyA tail. The dinucleotide anchor hybridizes to the last two bases of the mRNA. The oligo dT sequence 5' to the anchor hybridizes to the polyA tail from the first A of the tail and downstream toward the 3' end of the mRNA. For instance, the dinucleotide anchor of a first strand primer can have 12 different sequences, defined by any of A, C and G in the first position adjacent the oligo dT and any of A, C, G and T in the next position. Each different sequence can be used in a different anchor and by using the different anchors individually in separate reactions a population can be subdivided reproducibly into 12 different first strand cDNA sub-populations. Each sub-population being made up exclusively of individuals that t contain one of the dinucleotide sequences adjacent to polyA.

ANCHOR-INDEXING NUCLEOTIDES The phrase anchor-indexing nucleotides is used herein generally to refer to one or more nucleotides in an anchor that are used both for anchoring and for indexing. The number of nucleotides can be 1, 2, 3, 4, 5, 6, 7, 8 or more, as explained further elsewhere herein. For population profiling of mRNA or cDNA populations using strand displacement indexing adaptors in certain preferred embodiments of the invention in this regard, 2, 3 or 4 nucleotides are preferred.

ANCHOR INDEXING PRIMER The phrase anchor indexing primer is used herein generally to refer to a primer that comprises a portion for anchoring hybridization and a portion for indexing. The anchoring and indexing portions may be the same or they may be different. They may partially overlap or be separate.

ANCHOR-PRIMER A primer that comprises a region that acts as an anchor and a region that serves as a primer.

ARRAY The term array is used herein generally to refer to an arrangement in a fixed pattern. One type is illustrated by the 9 row by 12 column grid of wells in 96 well microtiter plates. Another type of array is illustrated by surface-immobilized arrays of cloned DNAs arranged in grids on the surfaces of membrane filters and by the surface-immobilized arrays of oligonucleotide probes attached in a grid on the surface of a glass "chip."

COGNATE The term cognate is used herein generally to refer to an indexing reagent, to its matching indexable characteristic(s), and to individuals in a population that possess that indexable characteristic and/or are recognized by the reagent. That is, as the term generally is used herein, an indexing reagent, the characteristic it specifically recognizes, and individuals in the population that possess that characteristic are cognates. The term has both theoretical and operational application; it is used to refer not only to theoretical matches but also to empirical outcomes of the interaction between an indexing reagent and individuals in a population. In an operational sense an indexing reagent and the individuals it recognizes reproducibly in a population are cognates, whether or not a single or multiple cognate indexable characteristics that are recognized by the indexing reagent exist, can be identified, or are unknown or known.

Put another way, Indexers, indexing reagents, indexing adaptors, parsers, parsing reagents, parsing adaptors and the like and the individual or individuals, specie or species, group or groups, type or types, sub-population or sub-populations and the like that they specifically and/or exclusively and/or reproducibly recognize generally are referred to herein as cognate and/or as cognates.

COMBINATORIAL The term combinatorial as used herein generally. refers to a structure made up of sub-units that can be combined in variety of ways. Frequently, but not necessarily, the structure is linear ad is made up of a string of the sub-units. Also frequently, but not necessarily, there are a limited number of well defined sub-units. A simple illustration of a combinatorial molecule is a stand of naturally occurring DNA, which is a linear string of nucleotides in which the nucleotide at each position is one of the four naturally occurring nucleotides (denoted by the letters A, C, G and T). Similarly, almost all naturally occurring proteins are linear strings of amino acids in which the amino acid at each position is one of the twenty naturally occurring amino acids. Glycols, such as glycols in glycoproteins, provide an example of branched combinatorial structures. Like polynucleotides and proteins, glycols are made of recurring sub-units (sugars). Glycols differ from one another in the number and arrangement of the sub-units. Unlike polynucleotides and proteins, however, the sub-units in glycols form branching structures rather then strictly linear ones. Other examples of combinatorial molecules are the very wide variety of product groups that can be made using combinatorial chemistry techniques.

COMPLETE SET OF INDEXING REAGENTS The phrase complete set of indexing reagents is used herein generally to refer to a set of indexing reagents of a given type that comprises indexing reagents with all the different indexing specificities possible for that type of reagent. Put another way, the term is used herein to refer to a set of indexing reagents of a given type that comprises indexing reagents that can index all the indexable characteristics that can be indexed by that type of indexing reagent. Put yet another way, a complete set of indexing reagents of a given type can index the entire indexing space defined by that type of indexing reagent. For instance, a complete set of indexing reagents that index polynucleotides by a two base long indexing sequence would contain at least one indexing reagent for each of the 16 possible two base indexing sequences.

HALF SITE Half site as used herein generally is short for restriction enzyme half site.

HOMOGENEOUS, HOMOGENEOUS MEDIUM The term homogeneous and the phrase homogeneous medium are used herein generally to refer to a medium in which components of interest are uniformly dispersed. Common homogenous media are solutions in which, by definition, dissolved solutes are uniformly dispersed.

INDEPENDENTLY DETECTABLE The phrase independently detectable as used herein generally refers to methods in which two or more things can be detected in the presence of one another. For instance, independently detectable labels can be detected independently of one another in the same solution. Fluorescent dyes used for DNA sequencing are an example of independently detectable substances. Another example is provided by independently detectable mass tags for MS detection.

INDIVIDUALS The term individuals as used herein generally refers to the entities that make a population. The entities (and thus the individuals) can be anything that can make up a population. In certain preferred embodiments of the invention, the individuals typically are molecules or complexes of molecules. In certain highly preferred embodiments of the invention in this regard, for instance, the molecules are cDNAs in a cDNA population.

INDEX The term index as used herein generally refers to a process of detecting the absence or presence and/or the amount of a sub-population of individuals in a population. More specifically, indexing generally refers to a process of detecting the absence or presence and/or the amount of a sub-population of individuals in a population.

INDEXING ADAPTOR The term indexing adaptor as used herein generally refers to an indexing reagent, particularly an indexing reagent that comprises another moiety, such as a moiety that binds individuals in the population or a moiety that can be detected. An example is provided by strand displacement adaptors, which comprise an indexing sequence for indexing polynucleotides and, adjacent thereto, a restriction enzyme half site, and further from the indexing sequence a sequence for amplification.

INDEXING REAGENT The term indexing reagent as used herein gene .rally refers to a reagent that specifically recognizes an indexable characteristic in individuals in a population. Indexing reagents in accordance, with certain preferred embodiments of the invention bind to individuals that comprises the indexable characteristic but not to other individuals in a population. Typically, when a binding reagent that recognizes a given indexable characteristic is allowed to interact with a population under specific-binding-effective conditions in accordance with the invention, the reagent binds to and/or tags the characteristic and individuals that contain it, whereby their absence, their presence and/or their amount in the population can be determined.

INDEXING SEQUENCE The term indexing sequence as used herein generally refers to a sequence that recognizes its complementary sequence in individuals in a population of polynucleotides. An indexing sequence in this sense can be viewed as the specific recognition element in an indexing reagent designed to index polynucleotides. Similarly, the sequence recognized by the indexing sequence is the cognate indexable characteristic.

INDEXABLE CHARACTERISTIC The term indexable characteristic as used herein generally refers to a feature, property, structure or the like that distinguishes some individuals from other, in some sense different, individuals in a population.

INDEXING SPACE The term indexing space as used herein generally refers to the number of different sub-populations defined by a given set of indexing reagents or by a given indexing procedure. It represents the total number of sub-populations that can be distinguished in a population by the set of indexing reagents or by the procedure.

INHERENT INDEXING SPACE The term inherent indexing space as used herein generally refers to the total number of different sub-populations that can be distinguished by a given set of indexing reagent, in and of themselves. In a sense R is a count of the total number of different indexable characteristics that are recognized by the members of the set.

MULTIPLEX The term multiplex as used herein generally refers to carrying out and detecting the results of two or more indexing reactions in the same volume, without the need to separate the products of the indexing reactions from one another in order to perform the detection method or to assess the results. Multiplexing generally is carried out using a different one of a set of independently detectable reagents to monitor the progress or outcome of each different indexing reaction of a set of indexing reactions being detected in the presence of one another. Specific embodiments involving multiplexing are described in the illustrative examples.

PARSE The term parse as used herein generally means much the same as the term index.

PARSING REAGENT The term parsing reagents as used herein generally means much the same thing as the term indexing reagent.

PHYSICAL SEPARATION The phrase physical separation as used herein generally refers to a process that removes or isolates individuals of one sub-population from the other individuals in a population.

POORLY CHARACTERIZED POPULATIONS The phrase poorly characterized populations as used herein generally refers to populations made up of a multiplicity of individuals of different types and in which the nature of the individuals, in particular the nature of differences between the different types of individuals are only partly known, are poorly known or are not known.

POPULATION The term population as used herein generally refers to a multiplicity of individuals; i.e., two or more individuals. Generally, preferred embodiments of the invention relate to populations of molecules and molecular complexes.

PRESENCE The term presence is used herein generally to refer to the presence in any amount, or to the presence in an amount sufficient for detection. Generally is used in the sense opposite to that of the term absent. Where the term present is used it implicitly means not absent. .

PROFILE The term profile as used herein generally refers to a partial or full description of a population in terms of different sub-populations and, typically, the number of individuals in each sub-populations. A partial profile describes a population by less than all its sub-population. A complete profile describes a population as a list of all its sub-populations and how much, if any, of each sub-population it contains.

RESTRICTION ENZYME HALF SITE The term restriction enzyme half site as used herein generally refers to the portion of the cleavage site of a site specific Type 11 restriction endonuclease that remains on one or the other side of the cut after double-stranded DNA is cut at the site by the restriction enzyme. For most, but not all, Type 11 restriction enzymes the half sites are the same on each side of the cut.

SDI The abbreviation SDI as used herein refers to strand displacement indexing.

SIZE SEPARATION The term size separation as used herein generally refers to a process or result of physical separation of two or mores sub-population from one another, in which all the individuals in a given sub-population are the same size and the individuals that make up different sub-populations are different sizes. An example is separation of polynucleotides by gel electrophoresis, such as electrophoresis through sequencing slab or capillary gels, for example, to separate fragments for expression profiling.

SPECIFIC RECOGNITION The recognition by an indexing reagent of its indexable characteristic in individuals.

SINGLETON The term singleton as used herein generally refers to the product of an indexing reaction that results from a single species of individual in a population. For instance, an indexing reaction that indexes a single mRNA of defined sequence in a population of many other mRNAs of differing sequence provides a singleton result.

SPATIALLY ADDRESSABLE The phrase spatially addressable as used herein generally refers to a plurality of reactions arranged in an array. Generally each position in the array is a particular one of a set of reactions and results for a given reaction can be determined by measuring the reaction product at the appropriate point in the array. Two types of arrays are particularly discussed herein. (1) Surface immobilized arrays are arrays in which one or more reactants are immobilized on a surface in a checkerboard or other pattern, with different reactants at different positions. Examples include Affymetrix GeneChipS™ and IMAGE clone arrays from Research Genetics, Inc. (2) Container or well arrays in which a multiplicity of containers, such as wells, are arranged in a pattern and different reactions are carried out at different locations in the pattern. A simple example is an array of reactions in a 96 well microtiter plate.

SPECIES The term species as used herein generally refers to a plurality of individuals that have one or more features in common that distinguish them from other individuals not of the same species. The common feature may be an indexable characteristic, and a species thus may be defined as all the individuals in a population that comprise the indexable characteristic and/or are recognized by an indexing reagent specific therefor. The common feature also may be intrinsic to the individuals, and an indexing reagent may recognize a limited part of the feature that also is present in individuals not of the same species. In that case individuals recognized by an indexing reagent may include individuals of more than one species.

Polynucleotides provide a concrete example in this regard. A species of cDNA may be defined by an indexing reagent that specifically recognizes the three base 3' sequence ATA. The species thus defined consists of all the cDNAs with the sequence ATA at their 3' end. If the cDNA is made from a complex mixture of mRNAs, the species defined by the ATA sequence will include cDNAs with many different sequences upstream of the 3' end ATA sequence. Another species of cDNA might be defined as all those molecules having a given sequence in its entirety. That is, for the most part, as a given molecular species.

The term is used herein generally to refer to a sub-population of individuals sharing a common feature of interest. In preferred embodiments the term may refer to molecules with a given indexable characteristic or it may refer to molecules with a single general structure. The terms subspecies, species, type, sub-group, group, family and the like are used somewhat interchangeably in different contexts; but, in a given context generally are used together to indicate hierarchical grouping.

SUBSTANTIALLY The term substantially as used herein generally refers to a significant fraction. For instance, the term is used herein in disclosing certain preferred embodiments of the invention in reference to the number of different mRNA species defined by entire sequence that will be represented in the singletons produced by an indexing procedure. To determine every distinct mRNA sequence in a complex mixture from a source like a human cell probably cannot be done by any indexing method. Such a perfect description would require sequencing with near perfect accuracy. Nevertheless, it is possible to determine many, most or nearly all of the different cDNA sequences in such a population using an indexable sequence much smaller than the entire sequence of any given cDNA in the population; but, long enough to define an indexing space large enough to parse many, most or nearly all of the cDNAs into sub-populations that contain mostly, nearly all or entirely cDNAs of identical sequence. In this sense the indexing method for cDNAs may be said herein to index substantially all the cDNA species in the population (defined by full length sequence), substantially all the sub-populations defined by the indexing reagents may be said to contain a single species of cDNA (defined by full length sequence), and substantially all the cDNA species may be said to be represented as singleton products ; although, more than one cDNA species may be detected together by some individual indexing reagents, some of the sub-populations defined by indexing may contain two or more different cDNAs and some of the cDNAs may not be represented as singleton results in any of the reactions.

STRAND-DISPLACEMENT INDEXING The phrase stand-displacement indexing as used herein generally refers to a method for indexing populations of polynucleotides in which an indexing sequence in an indexing reagent specifically hybridizes to its complement in a cognate polynucleotide and displaces a short complementary portion of the opposite strand of the polynucleotide. Stand displacement indexing is explained further below and in detail in references cited elsewhere herein.

SUB-POPULATION The term sub-population as used herein generally refers to a plurality of individuals that have a common feature that differentiates them in some way from other individuals in a population of individuals. Often the term is used to refer to a plurality of individuals that have in a common a given indexable characteristic that is recognized by a cognate indexing reagent to distinguish individuals in this sub-population from other individuals in a population of individuals.

DESCRIPTION OF FIGURES

FIGURE 1 is a schematic drawing showing several approaches for ligation-mediated indexing of polynucleotides.

The Top Panel, "L Primer-directed (AFLP)," illustrates primer directed AFLP. In the illustration a polynucleotide, represented by a line, is ligated at each end to an indexing adaptor, represented by boxes. Each adaptor comprises a restriction enzyme half-site represented by open boxes and a region comprising sequences for PCR amplification represented by grey boxes on the left and black boxes on the right. The left adaptor recognizes a n Mse I half-site and the right adaptor recognizes an Eco RI half site. The amplification sequences in the two adaptors can be the same or different so that amplification requires one, two or more primers, depending on the sets of adaptors and the protocol employed. Polynucleotides in a population are indexed, or parsed, or the like, by selective PCR using primers that are complementary to the adaptor and contain indexing sequences that further index the restriction fragments by hybridizing specifically to the sequence adjacent to the restriction enzyme half site. Polynucleotides recognized by and ligated to adaptors on both ends will be exponentially amplified by PCR, those that are recognized and ligated only on one end will be linearly amplified. Those that are not recognized on either end will not be amplified. A population of polynucleotides can be profiled by detecting and/or quantifying the PCR products generated by Primer-Directed AFLP reactions using different adaptors. AFLP methods are described in Money *et al., Nuc. Acids Res.* 24: 2616-2617 (1996) and Vos *et al., Nuc. Acids Res.* 21: 4407-4414 (1995) each of which is incorporated herein by reference in their entireties as to the foregoing particularly in parts pertinent to methods for primer-directed AFLP analysis.

The Middle Panel, "II(A) Adapter-directed, Class 11S," illustrates adaptor-directed ligation-mediated indexing in which indexable end sequences are generated by cleavage with a Class IIS restriction enzyme. A double-stranded polynucleotide is represented by two parallel lines. A recognition site for Fok I is indicated in the polynucleotide by the grey box above the label "Fok L" The left end of the polynucleotide was generated by Fok I cleavage from the indicated recognition site. The adaptor ligated to the left end of the polynucleotide recognizes the end-most four bases expose d by . Fok 1. The indexing sequence in the left adaptor is indicated by the black box and the four complementary bases exposed by Fok I are indicated by "NNNN." The rest of the adaptor, containing amplification sequences, is represented by the grey boxes. The right end of the illustrated polynucleotide was generated by Eco RI cleavage and is ligated to an adaptor specific for Eco RI-generated ends. The Eco RI half site is indicated by the open boxes. The rest of the adaptor, containing amplification sequences, is represented by the black boxes. As described for "I" in the top panel, adaptor recognition at both ends leads to exponential amplification. Recognition at one end leads to linear amplification and recognition failure at both ends precludes amplification. Exponentially amplified species are greatly preponderant after amplification and, in essence, at the only species detected.

The Lower Panel, "H(B) Class II," shows ligation-mediated indexing using a strand-displacement indexing adaptor on the left end of a polynucleotide and a restriction enzyme half site-specific adaptor on the right end. The polynucleotide is represented by the pair of parallel lines, except for the displaced indexing sequence represented by "N'N'N'N'," and the sequence complementary thereto (the indexed sequence) represented by "NNNN." The adaptors are represented by boxes. The left adaptor contains an indexing sequence, represented by black, that recognizes the indexed sequence NNNN and displaces the complementary sequence NN'N'N'(the indexing sequence) in the opposite strand of the polynucleotide. It also contains a Sau3AI half-site represented by a light grey box and a region comprising a sequence for amplification by PCR represented by the dark grey boxes. As described for the approaches shown in the top and middle panels of this figure, populations are profiled by determining the reaction of the population with each member of a set of indexing reagents, each comprising one or more of strand-displacement indexing adaptors. The presence of a reaction product and its amount for each of the reactions provides a profile that can be compared with profiles generated for other populations in the same way.

FIGURE 2 shows the structure of a strand-displacement indexing adaptor and the displacement structure it forms when ligated to a restriction fragment with a cognate end-sequence. The figure specifically illustrates the adaptor end, fragment end and displacement structure for a 5' overhang produced by digestion with the Sau3AI restriction endonuclease. The 8 bases of the protruding single-stranded end of the adaptor consist of the 4 bases of the Sau3AI half-site (3'-CTAG-5') and a four base indexing sequence (3'-XXXX-5'). The 8 base single stranded region anneals to its complementary sequence ("cognate") at the 5' protruding end formed by one strand of the target fragment (5'-GATCYYYY ---- 3'). As a result, the four nucleotides at the 3' end of the opposite strand of the fragment (3'-XXXX --are displaced by the 4 base indexing sequence of the adaptor. The 3' end of the non-invading strand of the adaptor is immediately adjacent to the 5' end of the protruding strand of the fragment in the strand-displaced structure. The strand break between them can be sealed with ligase, as shown by the bold face "G-7 in the displacement structure. The only free 3' end in the displacement structure is the 3' end of the displaced strand of the restriction fragment. The invading strand of the adaptor, which is not covalently joined to the fragment, is thermally displaced (∆) during the first cycle of PCR. The displaced 3' end then reanneals to the opposite strand of the fragment and is elongated by primer extension >), regenerating the template for binding to PCR primers.

FIGURE 3 shows much the same structures as Figure 2, except that it illustrates the adaptor and displacement structure for a 3' overhang, in particular that produced by the NlaIII restriction endonuclease. The 8 base long single-stranded protruding end of the adaptor consists of the NlaIII half-site (5'-CTAG-3') and a four base indexing sequence (5'-XXXX-3'). The 8 base single stranded region anneals to its complementary sequence (3'-GTACYYYY ---- 5') in the restriction fragment. As a result the four bases at the recessed

5' end nucleotides of the opposite strand of the restriction fragment (5'-XXXX ---- T) are displaced. Ligase covalently joins the non-invading strand of the adaptor to the non-displaced strand of the restriction fragment, rejoining the sugar phosphate backbone at the bases "A-G" shown in bold. The invading strand of the adaptor for the 3' overhang can serve as the PCR primer for the first PCR cycle, and the initial extension seen for 5' ends in Figure 2 is not required.

FIGURE 4 shows an embodiment in which a Taqman™ Probe is used for real-time detection of the PCR amplification of restriction fragments selected by strand displacement indexing. The figure specifically illustrates an adaptor similar to that shown in Figure 2 except that: (a) it contains a five base long indexing sequence ("IS") rather than one that is four base long; and (b) it also contains an additional sequence between the primer sequence and the restriction site (RS) that can hybridize to a Taqman™ probe, such as those from PE Applied Biosystems as described in Higuchi *et al., Biotechnology 11:* 1026-1030 (1993). The Taqman™ probe contains a fluorescent reporter ("R") (such as the fluorescent dyes FAM, TET, JOE, HEX, available from and described in literature provided by PE Biosystems, and a quencher ("Q"), such as the fluorescent dye TAMRA, also available from and described in literature provided by PE Biosystems. The adaptor and Taqman™ probe can be designed to use other portions of the adaptor for hybridization. For instance, as illustrated by. the dotted line and circle in the Figure, the probe can be designed to hybridize to a portion of the adaptor including the restriction site ("RS"). After ligation of the adaptor to its cognate cDNA restriction fragment, the fragment is amplified by PCR using the illustrated primer in conjunction with 3' anchor primers. Exponential accumulation of amplified product is monitored by hybridization of the Taqman™ probe during each cycle of the PCR. Non-hybridized probe is not detected due to fluorescence quenching; but, if unligated adaptor is not removed prior to PCR a low background might be detected at late cycles due to linear amplification of the adaptor primer strand. Multiplexing can be carried out using two or three of the three primers ("1,2,3") and Taqman™ probes shown in the figure. For a given restriction enzyme, PCR can be carried out using a primer-probe pair *(i.e.,* primer1-R1, primer2-R2, primer 3-R3) in conjunction with a 3'-N2N1-oligo(dT)-heel anchor primer. Fragments also can be-amplified using a primer corresponding to the primer of the primer-probe pair used for PCR; *e.g.,* fragments amplified using primer1-R1 can be amplified using primer1.

FIGURE 5 shows a schematic diagram of an adaptor for carrying out highly multiplexed stand-displacement indexing using 64 different mass tags, one for each of the 64 possible three base long sequences that can be formed by the four bases that naturally occur in DNA or RNA; A, C, G and T or A, C, G and U-respectively.

The adaptor comprises, from left to right: (1) a double-stranded region comprising a forward primer sequence (denoted by boxes with fine cross hatch, bracketed underneath and labeled "64 unique FP sequences"); (2) a double-stranded region comprising a universal sequencing primer (denoted by boxes with black spots on white, bracketed underneath and labeled "universal sequencing primer"); (3) a single-stranded region comprising a restriction enzyme half site (box with white spots on black, bracketed underneath and labeled "RS"), and immediately adjacent to the restriction enzyme half site, (4) a single stranded three nucleotide long indexing sequence extending from the end of the half site to the 5' end of the adaptor (box without pattern, labeled "5" at right, bracketed underneath and labeled "3-ntd IS").

Each of the 64 possible different three base indexing sequences is associated with a different, unique, forward primer sequence to form 64 adaptors. Each of the 64 different forward primers is labeled with one of 64 different cleavable mass tags. All 64 different forward primers can be used together in a single PCR reaction, together with one of the 192 possible N4N3N2N1-oligo(dT)-heel "anchor" primers as the 3' PCR primer. The mass tags can be detected independently of one another, and the PCR products for all 64 forward primers in each reaction can be determined together in a single SQMS analysis. A complete profile for the 192 x 64 indexing space defined by the 64 forward PCR primers and the 192 reverse PCR primers can be carried out in just 192 reactions.

Fragments thus amplified can be sequenced using a universal sequencing primer to carry out extension reactions, as represented in the figure. The extension products can be labeled in the reactions using fluorescent dye-labeled primer or fluorescent dye-labeled dideoxynucleotide triphosphates, in accordance with protocols for use with automated sequencing devices, for instance. Alternatively, for increased specificity, 64 different primers, one specific to each of the 64 adaptors, can be used for sequencing, rather than a single universal Primer (denoted by asterisks in the figure). The PCR FP sequences, moreover, can serve as sequencing primers that can be multiplexed.

DESCRIPTION

POPULATION INDEXING AND PROFILING

Many populations of interest for profiling are made up of discrete constituents (also referred to herein as individuals, members, objects and the like) of several distinct types (also referred to herein as species, classes, types, groups, sub-groups, sub-populations and the like). Particular instances of such populations can be characterized as frequency distributions of the number of individuals of each species, class, type, group, sub-group, sub-population or the like in the population. Such frequency distributions and expression and representations derived from them are referred to herein as, among other things, profiles, population profiles and the like. Generating such profiles is referred to herein as population profiling, profiling and the like, among other things.

Using frequency distributions of this sort to characterize and profile populations poses several related challenges. The-first challenge is to detect and distinguish individuals of given sub-populations from other individuals in a mixed population. The second challenge is to quantify the number of individuals in each sub-population detected in the population. That is, to quantify the number of individuals in the population in each sub-population of interest. Any method that distinguishes one sub-population from another may be used for profiling. Many different measures of quantity also may be used, including but not limited to direct, absolute measures of abundance and relative measures of abundance, such as measures relative to one or more other sub-populations in a given population, measures relative to one or more standards, and combinations thereof.

Both the number of different species, classes, types, groups, sub-populations or the like that are distinguished from one another in a population may be a few, some or all the different species or classes or types or groups or sub-populations or the like that actually occur or that might occur in the population being profiled.

In a particularly preferred aspect, the present invention provides profiling methods in which each species of a given statistically determined fraction of all the species of a certain genus in a population are detected and/ or quantified independently of one another in a population. In especially highly preferred embodiments in this regard, a substantial fraction of all possible species of a given genus are detected and/or quantified independently of one another in a population. In preferred embodiments in this regard, 50% or more of the different species are detected and/or quantified. In particularly preferred embodiments in this regard, 70% or more of the different species are detected and or quantified in the population. In especially particularly preferred embodiments in this regard, 80% or more are detected and/or quantified. In highly especially preferred embodiments in this regard, 90% or more are detected and/or quantified. In still more highly preferred embodiments, 95% or more are detected and/quantified. Among some most highly preferred embodiments, 97% or more are detected and/or quantified. In other most highly preferred embodiments in this regard, all possible species or other sub-populations recognized by a given set of indexers are detected and/or quantified in the population.

INDEXING REAGENTS

Another aspect of the invention relates to reagents that specifically recognize a cognate indexable characteristic and can be used to determine in a population the absence, the presence and or the amount of individuals that comprise the cognate characteristic. Such reagents generally are referred to herein as indexing reagents, but also are called parsing reagents, tagging reagents, recognition reagents, identifying reagents and the like, and as indexers, parsers and the like, among others. They are used for indexing, categorizing, dividing, splitting, parsing, apportioning, sorting, identifying, analyzing and the like, individuals in a population into sub-populations. Indexing reagents may be used to parse individuals in a population into sub-populations defined at any useful hierarchical level, including but not limited to individuals of a given type, sub-group, group, species, sub-genus, genus, family, class or the like.

In general indexing reagents in this regard specifically recognize a feature (referred to herein as an indexable characteristic) common to some individuals in the population that distinguishes them from other individuals in the population. The individuals that possess the characteristic for a given indexing regent form the sub-sub-population measured by that reagent. A given indexing reagent and the indexable characteristic it recognizes, as well as the individuals possessing that characteristic, are referred to herein as cognates of one another. An indexing reagent can recognize a cognate indexable characteristic in a cognate individual by any mechanism that provides a degree of interaction and specificity useful for indexing. That is, it can be any mechanism or combination of mechanisms that provides an interaction between cognates that has the specificity and/or exclusivity and/or affinity and/or avidity necessary for it to be used to determine with useful precision and/or useful accuracy the absence, the presence and/or the amount of cognate individuals in a population. The mechanism by which a given indexing reagent specifically recognizes its cognate indexable characteristic in an individual can be any physical or chemical interaction or combination thereof, including but certainly not limited to covalent bonding, van der Waals forces, hydrophobic interactions, hydrophilic interactions, electrostatic interactions, and other non-covalent types of bonding, and or combinations thereof to name just few.

Likewise, an indexing-reagent for use in the present invention can be formed of or comprise any material or chemical structure effective for specific recognition of a desired characteristic. Whereas the mechanism and structure of indexing reagents is relatively unimportant, specific recognition by an indexing reagent importantly should be reproducible so that the indexing reagent recognizes individuals in a given sub-population recognized by a given reagent generally are recognized by that reagent in any suitable population to be characterized, parsed, profiled or the like.

For instance, certain. particularly preferred embodiments of the invention in this regard relate indexing reagents that parse polynucleotide into sub-populations based on polynucleotide sequence, especially reagents that parse polynucleotides into sub-populations defined by the sequence of a few bases at the 3' end, the 5' end, or both the 3' and the 5' end of the polynucleotides. Particularly preferred embodiments in this regard relate to indexing reagents that parse polynucleotides into sub-populations based on the sequence of the last 2, 3, 4, 5 or 6 bases at the at the 3' end, the 5' end or both the Tend and the 5' end of polynucleotides, especially in this regard the last 3, 4 or 5 bases at the 3' end.

As noted above, the mechanism, structure or substance of an indexing reagent is not as important as its reproducible recognition of a specific characteristic that is useful for parsing a population into sub-groups. Thus, an indexing reagent for parsing a polynucleotide population on the basis of sequences can be or be comprised of, for instance, any polynucleotide that can form a perfectly matched duplex with its complementary sequence in individuals in the polynucleotide population to be indexed. That is, an indexing reagent according to these preferred embodiments of the invention can be comprised of a region that recognizes a specific base sequence by standard hybridization between a probe sequence in the reagent and its complementary sequence in individuals in the population, including polynucleotides, peptide-nucleic acids ("PNAs"), polynucleotide sequence-specific polyamides, and other reagents that form double-stranded or triple-stranded structures, sequence-specifically or involving recognition of other indexable characteristics. Other recognition reagents that may be used in this regard include an aptamer, a nucleic acid binding protein, and an antibody, to name just a few.

The sequence of an indexing reagent that specifically hybridizes to a cognate polynucleotides generally is referred to herein as an indexing sequence. The indexable characteristic, that is, the cognate sequence in the polynucleotide, generally is referred to herein as the indexable sequence or as the indexed sequence.

Certain highly particularly preferred embodiments the invention relate to indexing reagents that are stand-displacement indexing adaptors of the type described in Guilfoyle et al., Nuc. Acids Res. 24: 1854-1858 (1997) and PCT/US98/04819 (International Publication Number WO 98/40518) which are incorporated herein by reference in their entireties as to the foregoing particularly in parts pertinent to strand displacement indexers and strand displacement indexing.

INDEXABLE CHARACTERISTIC

An indexable characteristic is a feature of individuals that is specifically-recognized by an indexing-reagent. Individuals comprising an indexable characteristic constitute a sub-population that can be specifically measured by an indexing reagent that specifically recognizes that indexable characteristic.

An indexing reagent and the indexable characteristic it specifically recognizes generally are referred to herein as cognates of one another. That is, as cognate indexing reagent and indexable characteristic. An indexing reagent and individuals that comprise its cognate indexable characteristic also generally are referred to herein as cognates.

An indexing reagent that specifically recognizes an indexable characteristic in accordance with preferred embodiments of the invention is used to specifically measure the absence, the presence and/or the amount of a sub-population in a population to be profiled. A given specifically measured sub-population in this regard is made up of a individuals characterized by the indexable characteristic specifically recognized by the indexing reagent.

SETS OF INDEXING REAGENTS

In certain preferred embodiments in this regard the invention provides a plurality of indexing reagents to detect and/or quantify a given plurality of cognate species in a population. Such pluralities of indexing reagents and the like are referred to herein as sets, collections, kits and the like, among others. A plurality of cognate species recognized by a given plurality of indexing reagents is referred to herein as a set of cognates, cognate set, or a set of cognate species, classes, types, groups or other sub-populations.

In preferred embodiments of the invention in this regard, the reactions with a population of each indexer of a set of indexers is measured independently of the other members of the set. Each of the indexers in such embodiments is specific for a different cognate such as a species, type, group, set of groups or other sub-population. The independent measurements in such preferred embodiments detect the absence or presence of members of cognates recognized by each indexer in the population. In further preferred embodiments the independent measurements count or otherwise quantify the number and/or amount of members of each cognate species in the population.

Preferred sets in this regard comprise indexing reagents that detect and/or quantify independently of one another a given statistically determined fraction of all species of a certain type in a population. Particularly preferred embodiments in this regard, as described immediately above, relate to sets of indexing reagents for detecting and/or quantifying 70%, or more, more preferably 80% or more, still more preferably 90% or more, still more preferably 95% or more, in some still more highly preferred embodiments 97% or more and in other still more highly preferred embodiments 100% of possible cognate species in a population of the type to be profiled.

In especially preferred embodiments of the invention, a population of indexing reagents is utilized that comprises an adequate number of indexing reagents to categorize and/or quantify every different individual member of a population. A preferred example is provided by indexing-reagents which are indexing sequences for polynucleotide populations. An adequate number of indexing sequences can be determined by using statistical methods which analyze the polynucleotide population and predict how many different polynucleotide sequences are present, providing guidance on the number and length of indexing sequences which must be used to specifically-recognize each different polynucleotide molecule.

INDEXING SPACES

The total number of species that, can be recognized by a given set of indexing reagents defines the indexing space for that set of reagents. Although the indexing space is a fixed number of species for any given set of indexing reagents, a population can be parsed two or more times independently by the same set of reagents, in effect defining a much larger indexing space. Generally, this is possible when the features distinguished by the indexing reagents generally occur independently of one another at more than one place in individuals, and different occurrences can be accessed independently of one another.

In certain preferred embodiments in this regard the set of indexing reagents define an indexing space large enough to parse a genus, group, sub-population or population into all its individual species. Particularly preferred in this regard are one or more sets of indexing reagents that can be used together, either simultaneously or in succession, to parse a population into all possible species of a given type chosen for profiling therein. Especially highly preferred in this regard are one or more sets of indexing reagents that can distinguish all possible species of given type that can occur in a population, when used in concert either simultaneously, in succession, tandemly or in a combination thereof.

PROCEDURALLY DEFINED INDEXING SPACES

The manner in which indexing reagents are applied to index a population can be used to limit or expand the indexing space available to parse and to profile a population, as briefly noted above. The ability to define the indexing space not only by reagent design but also by procedural methods greatly expands the profiling capacity and effectiveness of regent sets and provides important practical advantages, including efficiency and economy of reagent use, as discussed below.

The indexing space inherently defined by a given set of indexing reagents is the number of different indexable characteristics that the reagents can recognize independently of one another. This is illustrated by the simple example of a set of indexing reagents that index polynucleotides based on a four base sequence. Inherently, these reagents can only distinguish between polynucleotides that differ from one another in the four base sequence recognized by these particular reagents. Since there are exactly 256 possible sequences of the four bases that naturally occur in DNA, this set of adaptors inherently defines an indexing space of 256 different sub-populations: each sub-population characterized by the presence in each member, of an indexable four base long sequence cognate to one of the 256 different four base long indexing sequences.

Indexing procedures that use a given set of adaptors are not limited to their inherent indexing space, however. Rather, the inherent space simply defines the maximum number of sub-groups that the indexing reagent can parse in a single step. If the same reagent can be used to parse the same .population in several independent steps it can be used to define a larger indexing space. Generally, the size of the indexing space defined by a series of independent indexing steps using one or more sets of indexing reagents will be the product of the inherent indexing spaces of the successively applied sets.

For instance, one-end recognition-based parsing of a polynucleotide population with a set of 256 indexing reagents that recognize all 256 four base sequences will index a population into 256 sub-populations, each made up of substantially all of the individuals in the population that have one of the 256 four base sequences in the single parsed end. If an indexing event is required at both ends, instead of just one, the same set of 256 indexing reagents defines an effective indexing space of 65,536 different sub-populations: the indexing space of the 256 adaptors applied to one end multiplied by the same 256 adaptors applied independently to the other end. Put another Way, indexing twice with four bases in independent events is the same as indexing once with eight bases and thus defines an indexing space of all possible 8-mer DNA sequences: 65,536.

Indexing spaces thus are defined not only by the inherent or intrinsic indexing space of an indexing reagent but also by how and how often it is applied to repeatedly index the population, both in independent indexing steps and in indexing steps that are not completely independent of one another. The indexing space furthermore is not necessarily defined by one set of indexing reagents. Indeed, more frequently than not it will be defined by several different sets of indexing reagents used in concert, often applied in succession, to index a population to a desired degree.

Again taking polynucleotides as illustrative, an example in this regard is provided by a procedure for parsing a complex mixture of cDNAs substantially to singletons using a succession of indexing reagents. The first indexing reagents are the primers used to make the first strand of the cDNAs. These indexing reagents parse the cDNAs into twelve independent sub-populations each defined by one of the twelve possible dinucleotides that can occur at the 3' end of mRNAs immediately next to the polyA tail. Since the primers hybridize to the tail as well as the dinucleotide anchor they have the general sequence 5'-T_{18}-V-N-3', where V is any one of A, C or G, and N is any one of A, C, G or T. The oligo dT track precludes T from occurring in the first position, thus there are 12 different first strand primers. Assuming approximately 15,000 different species of mRNA in the population, the 12 sub-populations will average about 1,250 different species, now in the form of cDNA.

The use of dinucleotide anchor primers for first strand cDNA synthesis, which provides just one method for sub-dividing an mRNA population into different fractions, is well known, as described for instance in Prashar and Weissman, Proc. Nad. Acad. Sci. USA 93: 659-663 (1996) and U.S. patent No. 5,712,126 to Prashar and Weissman each of which is incorporated herein by reference in its entirety as to the foregoing particularly in parts pertinent to the use of anchor primers to parse mRNA populations into two or more sub-populations defined by mono-, di-, tri- or other length anchor sequences at the 3' ends of oligo dT or other first stand cDNA synthesis primers.

Each sub-population, after conversion to double-stranded form, can then be indexed by a second method. 3' end fragments produced by a restriction enzyme then exclusively also might be parsed using strand displacement indexing adaptors that contain a four base indexing sequence. Indexing once with a four base indexing sequence would parse each sub-population defined by the first strand anchor dinucleotides into 256 sub-populations (referred to here as sub2-populations) that will comprise an average of approximately 60 different species of cDNA. A further indexing step can be applied to parse the 3,072 resultant sub2- populations substantially to singletons. A variety of ways can be employed to do this. The sub2-populations can be further indexed at both ends by amplification primers with defined dinucleotide anchor pairs that lie beyond the previously indexed sequences at each end of the individual molecules.

Another method for parsing mRNA populations involves synthesizing 12 cDNA sub-populations using dinucleotide anchors, cleaving each of the sub-population with a bank of 20 or so different restrictions, and then resolving mostly singletons from each of the 360 or so cleavage reactions by size separation on sequencing gels is described by Prashar and Weissman, Proc. Natl. Acad. Sci. USA 93: 659-663 (1996) and U.S. patent No. 5,712,126 to Prashar and Weissman each of which is incorporated herein by reference in its entirety as to the foregoing particularly parts pertinent to the use of restriction enzymes to parse a population of polynucleotides, particularly double-stranded DNAs, especially cDNAs. In the Prashar and Weissman method (referred to as READS) the 3' end fragments produced by restriction enzyme cleavage are selectively amplified via PCR amplification sequences in the cDNA synthesis primers and in an adaptor that is ligated to the ends generated by each restriction enzyme.

SITE SPECIFIC RECOGNITION AND CLEAVAGE

Site specific recognition and cleavage can be used in a variety of ways for indexing in accordance with some embodiments of the invention. The type and method of action of recognition and/or cleavage in this regard is not limited to any particular type or mechanism. The recognition may be independent of cleavage or associated with cleavage. Cleavage, where it occurs, moreover, may be at a fixed location within a, recognition site, on one side or the other of a recognition site, or at a fixed distance from a recognition site, among others. Particularly preferred reagents of this type in certain preferred embodiments of the invention are those that reproducibly recognize or cleave a cognate recognition or cleavage site, or both, in individuals in a population to be profiled. In this sense, the cognate recognition or cleavage sites are indexable characteristics.

Site specific recognition and/or cleavage reagents are useful in the invention in this regard in a variety of ways. They are useful as indexing reagents per se to recognize cognate indexable characteristics for assay to produce profiles, for instance. They also can be used to expose indexable characteristics in individuals in a population in a reproducible manner that can be exploited for indexing and/or profiling in accordance with the invention.

In certain preferred embodiments of the invention relating to polynucleotide populations, for instance, particularly double-stranded DNAs, indexing is carried out by a procedure that involves cleavage with, one or more sequence-specific cleavage regents. Cleavage reagents that can be used in this regard include, among others, Type 11 restriction enzymes, Class IIS restriction enzymes, other nucleases, ribozymes, chemical cleavage reagents, abzymes, and the like. Among preferred reagents in this regard are those that cleave at a defined sequence, such as Type 11 restriction enzymes that cleave within their recognition sequence, and those that recognize a defined sequence but cleave sequence-independently at a site in defined relationship to the recognition site, such as Class IIS restriction endonucleases.

Among particularly preferred Type 11 restriction enzymes, especially for use in strand-displacement indexing methods, are those that recognize a defined recognition sequence and cleave asymmetrically across the recognition sequence to produce single-stranded tails. In certain preferred embodiments, particularly some relating to strand-displacement indexing and profiling, Type 11 restriction enzymes are preferred that both recognize a four base sequence and cleave the recognition sequence asymmetrically so as to produce a four base long single-stranded "overhanging" end on both sides of the cut site. Particularly preferred in this regard are DpnII, MboI, NlaIII, Sau3AI, Tsp5091 and Tail. Also preferred in this regard are Type 11 restriction enzymes that recognize a four nucleotide sequence and produce after cleavage two base long overhangs on both sides of the cut site. Particularly preferred enzymes that produce two base overhangs include AciI, BfaI, Hha 1, HinP1I, HpaII, MseI, MspI, and TaqI. The use of restriction enzymes in this regard, especially those preferred enzymes listed above, is especially preferred in strand displacement indexing methods to expose indexable sequences in individuals in a population.

Many other restriction enzymes, including Type 11 enzymes and Type IIS enzymes, can used in accordance with certain embodiments of the invention in this regard. A comprehensive list of restriction endonucleases useful in this regard is provided by Robert and Macelis, *Nucleic Acids Research H:* 338-350 (1998) and at the website http:// www.neb.com/ REBASE each of which are incorporated herein by reference in their entirety (the latter as of the date of submission of this application to the USPTO) as to the foregoing particularly in parts pertinent to restriction enzymes useful for indexing.

Also useful, often in much the same way as restriction enzymes, are other reagents that effectuate site and or sequence-specific cleavage of polynucleotides or other types of individuals to be indexed.

POPULATION CHARACTERIZATION

Embodiments of the invention are applicable to indexing populations about which a great deal is known, including populations of individuals whose complete structures have been determined and validated. However, certain aspects and preferred embodiments of invention relate to populations of less well characterized individuals including poorly characterized individuals that nonetheless can be indexed using a set of indexing reagents that recognize a set of cognate indexable characteristics that can be estimated a prior to effectively parse the poorly characterized population to a given desired level of resolution. This aspect of the invention is illustrated by reference to certain preferred embodiments of the invention relating to combinatorial populations, particularly polynucleotides. However, the invention is generally applicable in other embodiments to other types of poorly characterized populations.

COMBINATORIAL POPULATIONS

Certain very highly preferred embodiments in this regard relate to indexing reagents that distinguish a combinatorial characteristic that differs between individuals of the different spècies to be detected and/or quantified in the population. A particularly preferred example of a combinatorial characteristic is a polynucleotide sequence, especially relatively short sequences referred to herein as indexing, parsing, tagging or identifying sequences as described in greater detail below. Particularly preferred sets of indexing reagents in this regard comprise indexers with each different indexing sequence defined by the combinatorial characteristic of a given combinatorial condition.

Particularly preferred embodiments in this latter regard relate to combinatorial characteristics that can be described mathematically as an ordered series or sequence of length "L" of a number of discrete characters "N", where "L" and "N" are positive integers.

Especially preferred indexing sets in this latter regard comprise indexing reagents that recognize each different combinatorial characteristic defined by the linear series. In certain particularly preferred embodiments in this regard the total number of combinatorial characteristics theoretically is the total number of possible different series of N characters of length L, defined by the equation N . Preferred sets in this regard comprise indexing reagents for detecting and/or quantifying independently of one another, individuals in each of the NL different species in the indexing space defined by the NL combinatorial characteristics.

Combinatorial populations in accordance with aspects of preferred embodiments of the invention in this regard are illustrated by populations of polynucleotides such as mRNAs, cDNAs and genomic DNAs and fragments thereof.

The individuals in populations of single stranded, naturally occurring DNAs, for instance, are linear strings of the four bases that naturally occur in DNA: A, C, G and T. Double-stranded DNAs are the same as single stranded DNAs in this regard, but are made up of two complementary single strands joined by base-paring. Because the two strands are complementary, the sequence of one strand in a double-stranded DNA necessarily defines the sequence. of the other strand. The same is true for RNA-DNA hybrids and other duplex or higher multimer molecules formed by base pairing and/or hybridization.

Different species or sub-populations of polynucleotides in such populations can be defined by sequence, which can be as long as the individuals that make up the sub-population or species, or as short as necessary to distinguish the individuals of a particular species or sub-population from the other individuals in the population. The sequences can be described mathematically as noted above as an ordered series of the four bases (thus N is four) of length L (which will differ for each full length sequence). The number of possible sequences for a polynucleotide population in which the individuals have an average length L is 4^L, as noted above. For a population of average length 1500, such as population of full length cDNAs, this is a very large number: 4^{1500}. Much the same is true for many populations of polynucleotides of interest, such as mRNAs and genomic DNAs.

However, typical samples, such as mRNA or cDNA made from tissues, mixed cells, pure cells and the like, even of complex organisms, such as humans, contain only a limited subset of the total number of possible sequences implied by the equation N^L. Human cells, for instance generally contain about 15,000 different mRNA species, far less than the 4^{1500} sequences possible for a 1500 base long RNA. The disparity between the large number of possible sequences and the much smaller number of actual sequences is typical of polynucleotide populations found in nature. In consequence, relatively short indexing sequences can be used to parse most to polynucleotide populations to the species level to an arbitrarily high degree of statistical assurance.

The length of such an indexable sequence, and the cognate length of the indexing sequence, determines parsing ability, and the degree of parsing of a combinatorial population, such as a polynucleotide population, can be controlled by the adjusting the length of the indexing and the indexable sequence. The relationship between the indexing space defined by the indexing sequence and the expected number of species in a population to be profiled determines the statistical expectation that a given indexable length will provide a given fraction of species-specific indexing results.

For a brief discussion of some applicable statistical treatments see U.S. patent number 5,871,697 of Rothberg et al. which is incorporated herein by reference in its entirety, as to the foregoing particularly in parts pertinent to sampling statistics of the type discussed above.

POLYNUCLEOTIDE POPULATIONS

Among the preferred combinatorial populations for profiling according to the present invention are populations of polynucleotides. Practically any population of polynucleotides can be indexed in accordance with the invention. For instance, the invention is useful for indexing populations of RNAs, such as hnRNA, unprocessed RNA, mRNA, rRNA, mitochondrial RNA, organelle RNA, tRNA, catalytic RNA and/or mixtures thereof. It is use ful for indexing populations of cDNA derived from and/or representative of populations of any of the foregoing RNAs as well. It is useful for indexing other DNAs too, including but not limited to genomic DNAs, particularly fragments of genomic DNA, especially fragments produced by sequence-specific cleavage. Particularly preferred in this regard are populations of mRNA and populations of cDNAs representative of mRNAs.

Moreover, the invention is useful to index polynucleotides from practically any source, including but not limited to a single-cell organism, or a mixture of single-cell organisms, for example, a bacterium, a protist, a unicellular algae, a fungus or a mixture of one or more thereof; one or more cells of one or more types from a multicellular organism or of a mixture of multicellular organisms, including those derived from natural sources and those cultured under controlled conditions. It is useful in particular to index populations of polynucleotides form cells in culture, from a tissue or a mixture of tissues, from an organ or a mixture of organs, and the like.

HOMOGENEOUS INDEXING REACTIONS ARE AMONG THE MOST HIGHLY PREFERRED EMBODIMENTS OF THE INVENTION IN CERTAIN RESPECTS

Homogeneous indexing reactions are among the most highly preferred embodiments of the invention in certain respects. Homogeneous indexing reactions in this regard are those that can be carried out homogeneously dispersed in a medium; that is, those that can be carried out entirely in a medium in which the individuals of the population to be profiled are randomly and or isotropically and or homogeneously distributed. Among preferred homogeneous media in this regard are those in which the sub-populations of interest are distributed the same as one another throughout the medium, most especially those in which @11 the sub-populations are randomly and or isotropically and or homogeneously and or uniformly distributed in the medium, especially those in which the sub-populations are homogeneously dispersed in the medium. In some highly preferred embodiments in this regard the indexing reagents are homogeneously distributed in the medium. Particularly preferred in this regard are those in which both the reagents and the individuals are homogeneously distributed. In some of the most highly preferred embodiments in this regard the results of indexing reactions can be detected, determined, measured, quantified and or monitored without having to physically separate different sub-populations indexed in the same indexing reaction. Especially preferred in this regard are embodiments in which cognate species are detected and/or quantified by their reaction with indexing reagents in the indexing reactions and without necessity for size-dependent physical separation of cognate species from other species in population.

In one aspect in this regard, preferred embodiments of the invention relate to methods that can be carried out in solution, dispersion, suspension, thin layer, monolayer, and the like, in which, typically, the individuals of the population to be profiled are dispersed in the media randomly without bias among sub-populations. Among particularly preferred embodiments in this regard a re those in which profiling reactions can be carried out and the results determined entirely in solution. Especially preferred in this regard are embodiments in which results can be determined in containers similar to or the same as the containers in which the reactions are carried out, particularly embodiments in which results can be determined in situ in the reaction container, especially methods in which progress of reactions can be monitored in situ in the containers over their time course from start or before to finish point or after, either discretely at selected points between or continuously or a combination thereof.

In preferred embodiments in this regard indexing reactions can be carried out in practically any type of containers or receptacles such as wells, tubes, bags, pouches, holders and other such devices. The containers may be separate from one another or may be part of a unitary device or part of a modular device comprising several unitary devices each of which comprises one or more containers. Preferred containers in this regard include but are not limited to wells of microtiter plates, including, among others, 96-well, 384-well, 864-well, 1536-well and 9,600-well microtiter plates, such as those that are commercially available from, for example, Corning Nunc, Imaging Research, and Pierce to name just a few suppliers.

In another aspect, preferred embodiments of the invention relate to one or more devices that comprise one or more containers that comprise indexing reagents. Particularly preferred embodiments of the invention in this regard are those in which. indexing reagents are distributed into an array of receptacles arrayed in a one or more unitary or modular devices. In preferred embodiments in this regard indexing reagents are distributed into wells of microtiter plates, including but not limited to 96-well, 384-well, 864-well, 1536 well and 9,600 well microtiter plates, such as those that are commercially available from, for example, Corning Nunc, Imaging Research, and Pierce to name just a few suppliers.

In some of the especially highly preferred embodiments in this regard indexing reagents for profiling a population are distributed into containers for profiling a population. In particularly preferred embodiments in this regard the indexing reagents comprise a set of indexing reagents for indexing a population to a given degree of completion and statistical reliability. In especially preferred embodiments in this regard the indexing reagents are SDI indexing reactions. In a further aspect in this regard, in particularly preferred embodiments of the invention the set of indexing reagents is distributed into wells of microtiter plates, including but not limited to 96-well, 384-well, 864-well, 1536 well and 9,600 well microtiter plates, such as those that are commercially available from, for example, Corning Nunc, Imaging Research, and Pierce to name just a few suppliers. In further particularly preferred embodiments in this regard the

DETECTING INDEXING IN THE PRESENCE OF NON-INDEXED INDIVIDUALS

In a related aspect of the invention, preferred embodiments employ methods that do not require a physical separation of indexed from non-indexed individuals to determine results of indexing reactions. In another aspect of the invention in this regard, preferred embodiments are those in which, in particular, results of indexing reactions can be determined without using a size-separation procedure that utilizes size differences to separate multiple species indexed in the same reaction and/or to separate indexed individuals in a reaction from the individuals that were not indexed.

In according therewith, detecting and/or quantifying the amounts of species in a population in preferred embodiments is carried out in the presence of other species. In particularly preferred embodiments in this regard species in a population are detected and/or quantified in homogeneous media and or homogeneously dispersed in a medium. In further particularly preferred embodiments in this regard indexed species are detected and/or quantified without size-dependant physical separation from other indexed and non-indexed species in the population. In especially highly preferred embodiments in this regard species are detected and/or quantified in the presence of other species, in a homogeneous medium, without size-dependent physical separation from other indexed or non-indexed species to be detected or quantified in the population.

SINGLETON SIGNALS AND STATISTICAL REPRESENTATION

In another aspect of the invention preferred embodiments relate to methods in which indexing reactions give rise to singletons; i.e., indexing reactions that give rise to a signal that is exclusively representative of the presence, and preferably, the amount of a single species or other sub-population in the indexed population. Particularly preferred embodiments of the invention in this regard relate to indexing reagents and methods in which the individual indexing reactions have a given desired statistical probability of indexing a single species in a population, of giving rise to a singleton and/or of providing a signal indicative of the presence and/or the amount of a single species in an indexed population. Preferred embodiments in this regard further relate to embodiments in which indexing reactions are multiplexed and each separately detectable indexing reagents in a multiplexed reaction has a given desired statistical probability of indexing a single species and giving rise to a singleton and/or providing a signal indicative of the presence and/or the amount of a single species in an indexed population.

As discussed elsewhere in the disclosure, species and sub-populations in this regard can be any species, sub-group, group, type, population or the like of interest. Typically, individuals of the species or sub-population or the like have in common an indexable characteristic that is not found in other individuals of different species or sub-populations. This does not mean, however, that the individuals in a singleton necessary are all identical. To the contrary, singleton in the broadest sense means that the signal from the reaction is indicative of one of a plurality of sub-populations of interest. This notwithstanding, however, a variety of preferred embodiments of the invention relate to indexing populations of molecules and to methods in which singletons are indexing reactions in which the indexing reagent indexes a single molecular species, the indexed individuals in the reaction are chemically the same (insofar as the specificity limit of the indexing reagent provides) and the signal arising from the singleton reaction is indicative of the presence and/or the amount of the that chemical molecular entity in the indexed population.

A highly particularly preferred embodiment in this regard relates to indexing polynucleotide populations, particularly mRNAs and polynucleotides representative of mRNAs, such as cDNAs, and genomic DNAs, particularly genomic DNA fragments that contain SNPs or other structural markers of actual or potential diagnostic or prognostic value.

KINETIC DATE: PCR AND REAL TIME DETECTION

In still other aspects, preferred embodiments of the invention relate to methods in which the kinetics of reactions between indexing reagents and cognate individuals in a population is used to detect and/or to determine the presence or the absence and/or the amounts of individuals of sub-population of interest in the population. A variety of different kinetic measurements are useful in the invention in this regard. End-Point detection utilizes data collected at one time at the end of a reaction. It does not provide any information about the reaction at different times. Thus, it cannot provide information about reaction kinetics. . In a sense, end point detection is the zero point on the scale of data useful for determining reaction kinetics. Perfectly continuous detection provides the other extreme of the scale. An uninterrupted perfectly continuous signal that accurately represents that state of the reaction at all points in time from start to finish can provide a complete set of data for determining reaction kinetics (as to the measured signal or signals). However, much less data than that provided by a continuous signal can be not only sufficient for a given application but considerably more economical.. Accordingly, a very wide range of signal sampling and time-elated data sets can be used to provide kinetic data for detecting or determining the absence or the presence and/or the amount of cognate individuals in à population

In preferred embodiments of the invention in this such regard kinetic data can be used to provide more accurate and more reproducible estimates of the absence, the presence and/or the amounts of cognate individuals in a population than can be achieved by end point estimates. Among benefits of kinetic data-based determinations using continuous or discontinuous signal sampling methods over by end point-based methods is greater sensitivity, improved reliability, improved quantitative accuracy and better reproducibility. However, while continuous sampling provides the most detailed information and therefore can be more sensitive and/or more accurate and/or more precise and/or reproducible, discontinuous sampling methods usually can provide virtually all or, at worst, most of much of the benefits of continuous sampling - in substantially less time with concomitantly lower cost. Accordingly, continuous sample methods typically are preferred where the need. for detail is particularly acute and the former considerations prevail. Discontinuous methods are preferred where considerations of cost and speed outweigh the advantages provided by the ultimate in detail. These and other considerations relating to kinetic methods are illustrated by way of the following discussion of certain preferred embodiments of the invention relating to reagents and methods for polynucleotide populations.

In certain preferred embodiments the absence, the presence and/or the amount of indexed polynucleotides in a population is detected and/or quantified by a process that involves first specifically amplifying the indexed polynucleotides in each indexing reaction and then determining the absence, the presence and/or the quantity of PCR amplification product for each indexing reaction. The results for each reaction can be used to evaluate the absence, the presence and/or the amount of the cognate polynucleotides in the population for the indexing reagent or reagents in each indexing reaction. As described elsewhere herein, typical PCR end point results can be used in the invention in this regard to provide both qualitative and quantitative information about polynucleotides in populations. However, in certain preferred embodiments time course measurements of PCR product accumulation are used to determine the absence, the presence and/or to quantify the amount of a cognate represented by a given reaction.

Kinetic measurement of PCR product accumulation of this type generally can be classified into two categories: (1) Real-time PCR detection methods, which provide a detailed time course of accumulation of one or more PCR products, by measuring product accumulation after each temperature cycle over the course of the amplification procedure. (2) Semi-real-time PCR detection-procedures in which PCR product accumulation is measured after some, but not necessarily all, cycles. In preferred embodiments of the invention in this regard kinetic data form real time detection or semi real time detection is used to provide more accurate and more reproducible estimates of the absence, the presence and/or the amounts of cognate polynucleotide in the population than can be achieved by end point estimates. Among benefits of real-time or semi real-time methods over the results provided by end point methods are greater sensitivity, improved reliability, improved quantitative accuracy and better reproducibility. While complete real-time procedures provide more detail and, accordingly, can be more sensitive and/or more accurate and/or more precise and/or reproducible, semi real time methods often provide much of the benefits of real time methods in substantially less time, using less sophisticated equipment, at lower cost. Accordingly, real time methods are particularly preferred in the present invention where the former considerations prevail, whereas semi-real time methods are preferred where the latter outweigh the former.

Certain preferred embodiments in this regard relate to profiling a population of polynucleotides in a homogeneous medium. Particularly preferred embodiments in this regard relate to detecting and/or quantifying. different species of polynucleotides in a population of polynucleotides to be profiled without physically separating the detected and/or quantified species from other species in the populations on the basis of differences in size. In another aspect in these regards, especially preferred embodiments of the invention in this regard relate to profiling a population of polynucleotides representative of gene expression in a sample. Particularly preferred embodiments in this regard relate to profiling a population of polynucleotides in a homogeneous medium, especially using indexing -reagents and methods that do not require physically separating indexed polynucleotides of one sub-population indexed in a reaction from polynucleotides of another, different sub-population indexed in the same reaction and/or from other polynucleotides in the reaction that were not indexed. Very highly preferred embodiments of the invention in this regard relate to profiling methods using strand-displacement indexing adaptors with a label for real time monitoring of stand displacement indexing results, kinetic data for determining the absence, the presence or the amount of cognate polynucleotides in the population, homogeneous indexing reactions and detection of indexing in the presence of other polynucleotides, potentially either indexed or not indexed, without the need to separate different species of polynucleotides indexed in the same reaction from one another or from polynucleotides that were not indexed in the population. In these and other regards, taken separately or in any combination, preferred embodiments of the invention are those in which the per cent representation, per cent singletons, and statistically estimated reliability, among others, are as stated above.

In certain preferred embodiments of the invention in this regard product accumulation is determined after each amplification cycle and the resultant values are graphed as a time course. A value for the threshold cycle ("C_T") is determined from the curve and is used as the measure of target initially in the PCR reaction. Since the threshold cycle occurs at the beginning of product accumulation, but in the linear portion of each time course, it generally is not distorted by non-linear effects that typify PCR reactions, such as saturation.

In accordance with another aspect of the invention, in certain preferred embodiments amplification products are measured in real time or semi-real time using any of several products and techniques well known in the art, including but limited to reagents and methods for time-dependent monitoring that utilize radioactive and/or fluorescent tags. Among especially preferred quantization systems for real-time and semi-real time detection are the Taqman™ system (Perkin-Elmer,. Foster City, CA); the Sunrise™ system (Oncor, Inc., Gaithersburg, MID); and Molecular Beacons (Tyagi and Kramer, *Nature Biotechnology* 14:303-308 (1996)).

Real time monitoring of PCR reaction progress and the outcome of indexing reactions is illustrated further by way of the examples below.

MULTIPLEXING

Generally, multiplexing techniques provide ways to detect and often to quantify several things at once. Typically, multiplexing techniques provide labels that can be detected independently of one another even when they are mixed together. The number of labels that can be combined under the reaction conditions of interest and detected independently of one another generally determines the number of different reactions that can be multiplexed in a single reaction: most often one reaction per independently measurable label. However,,many factors can affect the number of multiplexing labels that can be profitably employed under particular circumstances.

Multiplexing techniques can be used in accordance with certain preferred embodiments of the invention to index two or more different sub-populations in a single indexing reaction. In particularly preferred embodiments in this regard indexing reactions are multiplexed by attaching a plurality of different indexing reagents to plurality of different detectable labels that can be detected independently of one another in the same indexing reaction mixture. The signal from each of the different labels in a reaction mixture in accordance with these particularly preferred embodiments is used to detect and/or quantify the reaction of the associated indexing reagents independently of the others. Multiplexing in these preferred embodiments typically is used to carry out in a single container two or more indexing reactions that otherwise would have to. be carried out separately each in their own container from one another in two or more container

The use of independently detectable labels in this regard rests on much the operative principle as single lane DNA sequencing methods, in which four base-specific chain termination reactions are carried out in a single tube using four base-specific termination reagents each comprising a fluorescent label that can be detected independently of the fluorescent labels attached to the other three terminators. All the extension products from the reaction are resolved by electrophoresis on a single gel lane. Since the extension - products for each base are differently colored they can be distinguished from one another and the results from the four multiplexed reactions resolved in one gel are essentially the same as the results obtained by carrying out the four reactions separately and resolving the four reaction products of separately in four different electrophoresis gel lanes.

Several multiplexed reactions typically can be carried out much the same as a single reaction that is not multiplexed, except for the additional reagents required by the combination of reactions. Multiplexing thus generally greatly reduces the amount of common reagents required for the reactions, results in considerable saving of disposables, such as tubes, pipets, microtiter plates and the like, decreases the number of liquid handling steps required for a given number of reactions, thus considerably shortening the time required to prepare, process and read reactions, and reducing wear on equipment, and it also generally greatly decreases read out time since it often is possible to detect and quantify the signal from several labels at the same time as one another in a single tube or well.

The advantages of multiplexing in these respects are illustrated by the following general comparison of expression profiling human mRNAs using the same indexing procedure not multiplexed in one case and multiplexed in the other. Generally, the cDNA made from typical mRNA of a human tissue will contain - 15,000 transcripts. As discussed in examples below, profiling the cDNA population by stand-displacement indexing using 12 different anchor sequences (each 2 bases long), 1,024 indexing reagents each with a different five base long indexing sequences, and three different restriction enzymes (to expose different sequences in the cDNA to indexing by the reagents) requires 36,864 indexing reactions, of which only 33% are expected to provide a signal from one or more indexed and amplified cDNAs under optimal PCR conditions. No cDNAs will be indexed by the indexing reactions in fully 67% of the reactions. As a result more than two thirds of the wells or containers used to carry out the indexing reactions, and two thirds of buffers, common reagents, labels and detection time are expended on reactions that do not produce useful signals.

Multiplexing can greatly remedy this deficiency and provide considerable advantages not only in economy but also in speed. For instance, if the indexing reactions for three different enzyme were combined and amplified together using multiplexing techniques, the entire set of indexing reactions for all three enzymes could be carried out in the same number of steps and containers required for one enzyme without multiplexing.

In the example discussed above, for instance, multiplexing reduces the number of reaction mixtures from the 36,864 required to analyze three enzyme sets individually to 12,288. In a related aspect, multiplexing increases gene coverage per reaction and per reaction set. For instance, in the foregoing example multiplexing increases the coverage for a single set of reactions from about 83% to > 98%, since three way enzyme multiplexing provides three enzyme coverage in the same number set of reactions as one enzyme coverage without multiplexing. Yet another example of multiplexing is that at least one species of polynucleotide will be indexed and amplified in a much greater fraction of the multiplexed reactions the same reactions carried out without multiplexing. In fact, with three way enzyme multiplexing described above virtually all of the multiplexed reactions would yield a indexed and amplified product whereas only about 33% of the reactions provide indexed and amplified product without indexing as represented in Table 1.

Taqman™ probes are used for multiplexing SDI adaptors in accordance with some preferred embodiments of the invention in this respect. For instance, multiplexing with three enzyme sets for SDI using Taqman™ probes in accordance with some preferred embodiments in this respect can be carried out using probes and indexing adapters of the general designs illustrated in Figure 4.

For example, by way of reference to the foregoing example of three way enzyme multiplexing, three sets of enzyme-specific adaptors can be designed together with three different probes to multiplex the indexing reactions for all three enzymes. In accordance with certain preferred embodiments in this respect probe specificity for each enzyme is conferred by a sequence of nucleotides at the end of the probe that are complementary to nucleotides in the enzyme overhang sequence. Figure 4 illustrates a four base sequence, and the Figure captions explains further details relating to probe design. In accordance with preferred embodiments in this respect as well, each enzyme-specific set of adaptors also contains a region having a sequence to which a probe hybridizes.

In preferred embodiments relating to multiplexing in this regard, the probe-hybridization sequence in each enzyme-specific set of indexing reagents in a multiplexed reaction differ from one another. The different sets of adaptors also may have one or more amplification sequences that may be the same or different amplification sequences or both. In particularly preferred embodiments in this regard different probes are used to detect and or quantify the results of the different sets of indexing reactions that are multiplexed together. For instance, three different probes would be used to detect independently of one another the products of indexing by each of the three different enzyme-specific SDI adaptors in each multiplexed reaction in the example above. The majority of each probe can be complementary to a region in the adaptor, such as a the spacer region between an enzyme overhang sequence and a primer sequence. Specificity for a given enzyme can be conferred by a portion of the probe complementary to the enzyme overhang sequence.

Each probe would also contain a different fluorescent reporter group. Enzyme-specific amplification would be accommodated by using a different primer pair for each enzyme set employed to eliminate primer competition effects.

Other, similar, detection systems that can be employed in this way, include for example, Molecular Beacons and the labels used in the Sunrise TM amplification detection system.

Mass tags for mass spectrometry detection also are useful in this regard, as described in greater detail in examples below.

Additional reagents and methods that can be used in the invention in this regard are available and likely will be developed in the future. As more independently detectable labels become available and greater multiplexing becomes practical, the number of transport and liquid handling steps can be reduced by their use in certain preferred embodiments of the invention.

Aspects of the invention are illustrated by the following examples. The examples are exemplary of the invention. They describe specific embodiments of the invention and they do not relate its limitations.

EXAMPLE 1

M deling Ind xing Procedur s by Databas Analys s: Human C ll Expr ssion Profiling Mod ls

(A) EGAD Transcript Database Analyses

A database of human transcript sequences was used as a model system to predict results of indexing procedures for profiling gene expression in human cells. The predictions were generated using the EGAD database, which contained 7216 unique transcript sequences at the time of the analysis. The database was analyzed to identify all the 3' end fragments generated by cutting the full length sequences in the transcripts at all the sites for cleavage by the each of the restriction enzymes Sau3AI, Tsp5091, Tail and NlaIII. The 3' fragments for each enzyme thus identified were further categorized by their N 1 N2 anchor sequences at their 3' end just 5' to the polyA tail, and then by the 1024 5 base long indexable sequences at their 5' ends.

Three of the four enzymes, the 12 anchor sequences and the 1024 5 base indexing sequences together defined an indexing space of 36,864 signatures. All four of the enzymes, the 12 anchors and the 1024 5 base indexing sequences together defined an indexing space of 49,152 signature. The number of matching 3' fragments identified in the database was counted for each of the 36,864 signatures defined by the three enzyme procedure and for each of the 49,152 signatures defined by the 4 enzyme procedure.

The results are summarized in Tables 1 and 2.

The results for a three enzyme analysis using Sau3AI, Tsp5091 and Tail, but not NlaIII is summarized in Table 1. The table gives a count of the number of signatures that do not match any fragments in the database ("0") and that match 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 and more than 10 fragments (denoted by rows labeled accordingly under "frag no." along the left side of the table). The table breaks the data down by dinucleotide anchor as well (listed at the head of each column).

For instance, 2,280 of the signatures containing the dinucleotide anchor "AC" did not match any of the 3' fragments produce from the database by any of the three enzymes. 600 signatures matched 1 fragment produced by the three enzymes, 141 signatures matched two fragments, 33 signatures matched three fragments, and so on. In essence the table shows how many fragments each 5' indexer/3' N1N2 - signature would detect in an experiment carried out according to the precepts of the database analysis. It thus provides useful guidance both for designing indexing procedures and for assessing the actual results of profiling experiments.

For three enzymes, as summarized in Table 1, 24,861 (67%) of the 36,864 signatures did not match any of the fragments derived from the database. 8,141 (22%) matched one fragment, and would have provided singleton signals indicative of the representation in a population of the cognate mRNA. 3,862 (11%) of the signatures matched two or-more fragments. Overall, there would have been 0.492 cognate fragments for each signature. This results was extrapolated to yield an average of 1.02 cognate fragments per signature in a population of 15,000 transcripts. This indicates that the indexing procedure, on average, would result in PCR amplification of a single species in a significant number of wells.

This point is underscored by considering the fraction of signals that arise from indexing a single cognate species as opposed to multiple species. 67%of the 36,864 separately detectable indexing reactions that would be carried out to index a population according to this scheme would not have produced a signal. Of the remaining reactions, which do produced a signal arising from one or from multiple cognate species, 68% would have arisen from indexing a single cognate species.

Table 2 shows the results of the same analysis carried out with all four enzymes mentioned above. The results are much the same as those obtained with three enzymes. 67% of the complete set of 49,152 signatures define by the four enzyme procedure did not match any fragments from the database, and would not be expected give rise to a PCR signal. The remaining 33% of the signatures matched one or more cognate species from the database and would be expected to give rise to a signal. Of these, 22% of all signatures; but, 66% of the signatures that give rise to a signal, would be derived from singleton indexing reactions.

Further statistical evaluation showed that the three enzymes, each recognizing a different four base long sequence, together with 12 dinucleotide anchor sequences and 1024 five base indexing sequences provide 98-99% gene coverage (taking into account overlaps of fragments detected as singletons by more than one enzyme). Two enzymes provides 96% coverage. With one enzyme the coverage drops to 83%. Increasing the number of enzymes to four provides only a modest boost in coverage to over 99%.

The results obtained by these database analyses are used to benchmark results that were obtained using fluorescent and Mass Tag probes in the following examples.

(B) 3' Index Database Analyses

Similar analyses were carried out on an expanded scale using a database of 3' EST sequences derived from the GenBank dbEST database. The database, called the 3' Index Database, contained about 40,000 ESTs from the GenBank database. Approximately 20,000 3' Index ESTs highly validated for the sequence of the four anchor nucleotides immediately upstream of the polyA tracts were used in the analyses. The larger size of the database more closely approximated the population diversity expected for human cell mRNA populations. The high reliability of the four base sequence at the 3' end of the mRNA sequence increased the accuracy of the model for indexing these sequences and provided increased confidence in comparing the computer generated results with those obtained using Mass Tags, described in Example 5. below.

The scheme for indexing for this analyses selected 3' end fragments generated by three enzymes and parsed the fragments by a 3 base indexing sequence on the restriction cut end and a 4 base anchor-indexing sequence on the 3' end. The three enzymes, 64 sequences of the 5' end 3 base indexing sequence and the 192 sequences of the 3' end 4 base anchor indexing sequences together defined 36,864 unique signatures: (3)(64)(192) = 36,8,64. For each enzyme taken alone, the 5' and 3' indexing sequences together defined an indexing sub-space of (64)(192) = 12,288 unique signatures.

An analysis and results for Sau3AI taken alone generally -illustrates the results obtained using this database. Sau3AI 3' end fragments less than 750 base pairs long were identified in the database. The 3 indexable sequences at the 5' end and the 4 indexable sequences at the 3' end of each of the Sau3AI 3' end fragments thus identified then were matched to with cognate signatures. Finally, the number of cognate fragments was counted for each signature. As expected (and as described above) most of the signatures did not match any of the fragments identified in the database. Of those that did, 2,437 were singletons, 767 matched two different cognates in the database, 271 matched 3 cognates, and 130 matched more than 3. Thus, nearly 70% (67.6%) of the signatures that matched a cognate in the database matched one unique cognate. Therefore, nearly 70% of the signals from a profiling experiment according to this scheme will provide information about a single species of mRNA in a population of mRNAs about as diverse as the mRNA in a human cell.

The same analysis carried out for: Tsp5091, Tail and NlaIII yield much the same results.

EXAMPLE 2

Automated Homogeneous Indexing: SDI Expression Profiling (unmultiplexed)

mRNA is extracted from a human cell line according to conventional methods. cDNA molecules corresponding to mRNA molecules in the nucleic acid sample are synthesized according to the Gubler-Hoffman method. First strand synthesis of the cDNA is carried out using oligo-dT primers with dinucleotide 3' anchors and 5' heels comprising amplification sequences. Twelve different primers are used, one for each of the 12 different dinucleotide anchor sequences. Three different heel sequences are used in the 12 anchors, a different one for each of the three N1 nucleotides adjacent to the oligo(dT) sequence. Twelve separate ONA synthesis reactions are carried out, one for each of the 12 different dinucleotide anchor primers. In essence, the 12 different dinucleotide anchors in the first strand primers index the mRNA population into 12 different sub-populations of cDNA, based on the sequence of the last two bases of the mRNAs, just before the polyA tails.

Each of the twelve cDNA sub-populations then is separately digested with the restriction endonucleases Sau3AI, Tsp 5091 and Tail, 36 digestions in all. Each of these enzymes has a 4-nucleotide recognition sequence and generates a four nucleotide long single-stranded tail (also referred to as a cohesive end) on both sides of the cleavage site. Reactions are carried out using standard procedures according to manufacturers recommendations.

The products of the restriction digests are further indexed by strand displacement indexing ("SDI") using three sets of strand displacement indexing adaptors with 5 nucleotide long indexing sequences. Each set of indexing adaptors has a 4 nucleotide sequence complementary to one of the tails produced by one of the three enzymes used to digest the cDNA. The adaptors also contain a sequence for amplification by PCR. Structures of SDI adaptors and indexing products with [tetranucleotide indexing sequences] are illustrated in Figures 2 and 3. Each set of indexing adaptors contains .1024 adaptors, one for each of the 1,024 possible 5 base long sequences of the bases A, C, G and T.

The three sets together contain a total of 3,072 different SDI adapters. All 3,072 adaptors are used to index each of the 12 sub-populations of cDNA. The procedure thus defines an indexing space of 36,864 unique signature that can parse the mRNA populations into 36,864 unique sub-populations. Each signature is defined by a unique combination of (1) one of the 12 dinucleotide anchors sequences, (2) one of the three restriction enzyme cleavage sites, and (3) one of the 1,024 five base long indexed sequences. Each sub-population is made up of the mRNAs that match one of the 36,864 unique signatures. The indexing space of 36,864 unique signatures is more than twice the number of 15,000 different species of mRNA estimated to be present, on average, in a human cell. Computer modeling using EGAD shows that an arbitrarily high fraction of the indexing reactions will index a single species of mRNA in the foregoing procedure, using just a few enzymes. For instance, using just two enzymes in the foregoing procedure will provide coverage of about 83%, of the unique signatures in a typical human mRNA population, according to computer modeling using EGAD. The analysis also shows that using three enzymes will provide about 96% coverage , and that using four enzymes will provide more than 99% coverage.

Indexing adaptors are robotically distributed into wells of microtiter plates, 36,864 adaptors in all. Each adaptor is placed in a different well, 36,864 wells in all. The adaptors are arranged in 12 banks (or sets), one for each of the 12 differently anchored cDNAs. The adaptors in each bank are arranged in three sub-banks, one for each of the three different restriction enzymes reactions used to cut each of the 12 differently anchored cDNAs. Accordingly, each sub-bank is made up of 1,024 wells, each containing a different one of the 1,024 adaptors defined by the 1,024 different 5 base long indexing sequences .

The 36 restriction digests then are robotically aliquoted into wells with the adaptors so that each of the sub-banks is mixed with a different anchor-restriction enzyme combination. SDI reagents are added and the mixtures are adjusted to conditions effective for SDI and ligation and then incubated.

SDI reactions and ligations are carried out on the restricted cDNA as described PCT/US98/04819, International Publication Number WO 98/40518 and in Prashar and Weissman, *Proc. Natl. Acad. Sci. USA* 93: 659-663 (1996) and U.S. patent No. 5,712,126 to Prashar and Weissman, which are incorporated herein by reference in their entireties, as to the foregoing particularly in parts pertinent to carrying out the indexing and ligation reactions).

EXAMPLE 3

Real Time Detection:

Taqman™ SDI Expression Profiling

Following the SDI reactions, indexed Tend fragments in each well are amplified by PCR. In this example a single forward and single reverse primer is used for all of the PCR reactions. The forward reactions has the sequence defined by the sequence for amplification in the SDI adaptors used for indexing. The reverse primer has the sequence defined by the sequence for amplification in the heel of all of the primers used for the 12 first strand cDNA synthesis reactions. To confer greater specificity in the PCR step, the reverse primers can be the entire dinucleotide-anchored primers used for cDNA synthesis.

PCR reactions are monitored in real time using Taqman™ assays. All the reactions are carried out using universal Taqman™ probes. The general structure of the probes is much like that shown in Figure 4. The probes that are used in this example do not hybridize to either the indexing sequence or to the restriction enzyme half site, so that the same Taqman™ probe can be used in all of the reactions. Taqman™ probes also can be designed to hybridize to part or all of the restriction half site and or indexing sequences in an adaptor, a feature that can be used to confer additional specificity. The Taqman™ probe that is used in this example is obtained from Perkin-Elmer. It has a very low detection limit and provides very high sensitivity detection. The probe also provides useful quantitative results over a broad dynamic range and can be used to detect differences as small as two-fold in comparable signals.

The PCR reactions are carried out according to the instructions for Taqman™ assays provided by Perkin-Elmer. Fluorescence of the probes is read on a PE 7700 (Perkin-Elmer, Applied Biosystems, Foster City, CA). Probe fluorescence is determined after each cycle of the PCR reaction. The number of cycles analyzed for each reaction, the selection of the threshold cycle (C.,) and the generation of standard curves is determined in accordance with the manufacturers recommendations.

EXAMPLE 4

Multiplexed Real Time Indexing
Multi Fluorescence SDI Expression Profiling

Strand-displacement indexing with real time detection is carried out as described in Examples 2 and 3 to profile whole genome expression in mRNA from a human cell line. The cell line, mRNA extraction and purification, cDNA synthesis primers and synthesis, restriction digests, indexing reagents, indexing and ligations reactions, PCR primers, PCR amplification procedures, Taqman™ probes and detection methods and instrumentation all generally are the same as described in Examples 2 and 3, except for the differences explained below.

Instead of one label, as in the previous examples, three different fluorescent dyes are used that can be detected independently of one another in real time. A different dye is used to assay results for each of the three restriction digests, so that the digests can be assayed together. Instead of the three banks of wells described in Example 3, only a single bank of wells is used.

Each well contains three indexing adaptors instead of one - one for each restriction digest. For convenience, the three adaptors in each well in this example have the same indexing sequence. And, instead of distributing each digest into a different bank of wells, all three digests are introduced into the same bank, thus reducing well, liquid and plate handling requirements by two thirds.

Instead of a single Taqman™ probe, as in Examples 2 and 3, three different probes are used. Each of the three probes is labeled with one of the three fluorescent tags. Furthermore, unlike the single probe in Examples 2 and 3, which hybridizes to the same sequence in all of the indexing adaptors, each of the three different probes in this example hybridizes specifically to the indexing adaptors for one of the three enzymes, and not to the adaptors for the other enzymes. The general structure of the adaptors in illustrated in Figure 4. As shown in the figure, each probe hybridizes specifically to the restriction enzyme specific portion of an adaptor; each probe thus is specific for fragment ends produced by a specific restriction enzyme ligated to a restriction enzyme -specific adaptor. Each of the 'differently labeled probes thereby can be hybridized exclusively to the indexed fragments from a specific restriction digest, and the signal from each label thus provides the read out for one of the restriction digests, but not the others.

The instrument for reading PCR results is able to read the fluorescence from all three labels in a single well independently of one another, with little time penalty over that required to read a single label. Otherwise, except for cross-talk corrections (where appropriate) data collection and analysis is the same for each dye in this example as for the single tag used in Examples 2 and 3.

Results obtained using the multiplex approach are comparable to those obtained using a single tag.

EXAMPLE 5

High Throughput Profiling
Based On Highly Multiplexed Homogeneous Indexing:
SDI Expression Profiling Using Mass Tags

The following example illustrates ways in which independently detectable indexing reagents can be used together in a single indexing reaction to index fragments produced by a single restriction enzyme. This multiplexing approach contrasts with examples in which multiplexing is used to index fragments produced by different restriction enzymes. The example also illustrates the use of mass tags for multiplexing, in contrast to multiplexing using fluorescent labels for detection in other examples.

mRNA is extracted as described in Example 2. cDNA is prepared from the mRNA either in (a) a single reaction containing a mixture of anchor primers or (b) twelve separate reactions individually containing one dinucleotide anchor primer, as described in the foregoing example. The cDNA is digested with Sau3AI generating a four nucleotide long single stranded tail on both. sides of the cleavage site. Most fragments that are produced by the Sau3AI cleavage have Sau3AI half sites on both ends. Fragments that contain the 5' end or the 3' end of the a cDNA have a Sau3AI half site only on one end. And c ' DNAs that are not cut by the enzyme do hot have a Sau3AI half site on either end. Results much the same as those described for Sau3AI also are obtained using the restriction enzymes Tsp5091 and Tail, which also recognize a four base pair sequence and produce a single-stranded tail that facilitates ligation.

The population of cDNA fragments produced by the Sau3AI digestion are indexed using a set of 64 Sau3AI-specific indexing adaptors having 3 base indexing sequences (referrer to below as "3N" or "3N-associated" adaptors). The adaptors each comprise a 7 base long single stranded overhang. The end of the overhang is a three base indexing sequence. The four base long Sau3AI half site makes up the rest of the overhang. 64 different adaptors are employed, each one having one of the 64 different 3 base sequences that can be formed by the bases A, C, G and T. The adaptors also differ in amplification sequence so that each indexing sequence is uniquely associated its own amplification sequence, different from the others. cDNA fragments indexed by given adaptor therefore can be amplified independently of other cDNA fragment using adaptor-specific primers defined by the adaptor-specific amplification sequence. In addition, the primer sequence for amplifying a given adaptor can be uniquely labeled with a mass tag different for that of all the others. Thus, the PCR products formed by each adaptor-indexing sequence are associated exclusively with a unique mass tag. Finally, the adaptors can also comprise a nested primer sequence for sequencing. The sequence can be "universal," so that one sequencing primer can be used with all adaptors. Alternatively, different adaptors may have different sequencing primer sequences, so that they can be differentiated from one another and multiplexed. For instance, as described above, a set of 64 different adaptors differentiated by 64 different indexing sequences and by 64 different mass tags, also may be differentiated by 64 different sequencing primer sequences. Then, products incorporating any one of the 64 different adaptors can be sequenced independently of the others, even when indexing is carried out using all 64 adaptors together in the indexing reaction.

An aliquot of the Sau3AI digested cDNA and is mixed with all 64 indexing adaptors in a single reaction vessel. The mixture is incubated under conditions for strand displacement indexing and for ligation, so that the non-displacing strands of adaptors are ligated to the non-displaced strands of cognate polynucleotides to which they are hybridized.

The ligation products then are further indexed, using a set of 192 reverse primers with 4 base long indexing sequences (referred to below as "4N" reverse primers). The primers each comprise, in the 5' to 3' direction, an oligo dT tract immediately followed by a 4 base long indexing sequence. Each of the 192 primers comprises one of the 192 different 4 base long sequences that can be formed by the four bases A, C, G and T. The 192 reverse primers hybridize only to cdna fragments that contain a poly dA tract. They hybridize for the most part only to the poly dA tract at the 3' ends of the 3' end fragments generated by Sau3AI cleavage. As a result, they support polymerization only in the reverse direction form the intact 3' ends back toward a Sau3AI cut (or the 5' end of an uncut cDNA). They thus support amplification only of the 3' end fragments. Furthermore, they support exponential amplification of the fragments only in the presence of other primers that can primer the forward reaction.

Exponential amplification by PCR Is carried out using the 4N oligo dT reverse primers together with a set of 64 N3N2N1 ("3N-associated") forward primers. The 64 forward primers match the amplification and indexing sequences of the 64 3N-associated Sau3AI adaptors used for the first indexing step above. The primers are designed to have similar melting temperatures so that they all can be used with the same set of thermal cycle conditions. Most of the amplicons from the PCR reaction result from reverse priming by a 4N primer on the 3' end and by a 3N-associated primer on the 5' end. However, a small fraction of the amplicons result from priming on both ends by a forward primer. The fraction can be reduced, if necessary or desirable, by using longer indexing sequences or by additional indexing steps.

The 64 forward primers and the 192 4N reverse primers as used in the forgoing procedure define an indexing space of 12,288 different sequence signatures. This is about the number of different mRNAs expected in a typical human cell (generally estimated to be about 15,000). A substantial fraction of the 12,288 indexing reactions thus should give singles that arises from a single species of mRNA, that is, from a singleton.

Each different forward primer is tagged with a different mass tag (Masscode™ tags from Rapigene™). The tags differ by 4 atomic mass units and can be detected and quantified independently of one another by single quadrupole mass spectrometry (herein abbreviated "SQMS"). The tags are attached to the primers by a photosensitive link. The tags are employed as described by the manufacturer (Rapigene™, Bothell, Washington, USA).

Indexing, tagging and amplification are carried out using the reverse primers and the mass tagged forward primers as follows. The 192 4N-oligo dT reverse primers are introduced into 192 wells in two 96 well microtiter plates, a different primer in each well. A mixture containing all 64 different mass-tagged forward primers also is added to each well. Aliquots of the ligation reaction are added to each well and, following addition of PCR reagents, the mixtures are subjected to PCR amplification.

The products of the PCR reactions are subjected to SQMS to determine how much of each mass tag has accumulated in each well and, thereby, how much of each PCR product accumulated in the reaction and how much of each cognate cDNA, and mRNA from which it was derived, there is in the sample. Unincorporated primers, which may generate noise and or other spurious signals is reduced or eliminated by standard techniques, such as binding to ion exchangers and or dialysis, among others, adapted to microtiter plate formats and automation. The PCR products in each well then are subjected to photolysis to release mass tags. A small fraction of each reaction is analyzed by SQMS, according to the manufacturers instructions. The mass spectrum for each well directly provides the quantity of product accumulated for each of the 64 different mass-tagged forward primers and one of the 192 reverse primers. 192 samples are analyzed, one from each well.

Essentially all steps are carried out robotically using widely available equipment for handling and using microtiter plates. SQMS analysis is carried out using automated equipment and instrumentation in accordance with the manufacturer's recommendations. The entire analysis is completed in approximately one day: roughly 10 seconds per mass tag, 11 minutes per well and about 30 hours for about 200 wells and more than 12,000 signatures.

The foregoing illustrative embodiments provide various examplars of the invention that do not depict or otherwise indicate its limitations. Rather the foregoing description of the invention is provided to enable those skilled in the pertinent arts to ascertain the general spirit and scope of the invention, and to make and use the invention. It is clear that those of skill will be enabled to elaborate on the disclosure to make various changes and modifications of the particularly disclosed embodiments that nonetheless do not depart from the spirt and scope of the invention, and to adapt it to various usages and conditions.

Incorporation by-reference

The entire disclosure of all applications, patents and publications, cited above and in the figures are hereby incorporated by reference.

Cross-reference to related applications

This provisional application relates-in-part to International Application Number PCT/US98/04819 (International Publication Number WO 98/40518) designating the United States, and of U.S. Application Number 08/815,448 filed 11 March 1997, both of which are herein incorporated by reference in their entireties. It is intended that this provisional application when perfected as a standard utility application will be filed as continuation in part of both above-mentioned applications: PCT/US98/04819 (International Publication Number WO 98/40518) U.S. Application Number 08/815,448.

TABLE 1

EGAD 7216 TRANSCRIPTS
12 DIFFERENT 2-BASE ANCHOR SEQUENCES
1,024 DIFFERENT 5-BASE INDEXING SEQUENCES
3 DIFFERENT ENZYMES: Sau3AI, Tsp5091, Tail

frag #	AC	AG	AT	CC	CG	CT	GC	GG	GT	TC	TG	TT
0	228	194	207	193	263	218	225	217	232	162	172	169
	0	8	8	8	0	5	6	2	7	5	7	5
1	600	748	695	758	374	656	610	642	582	838	823	815
2	141	251	192	260	58	172	154	181	126	356	310	334
3	33	86	68	70	6	39	36	43	26	150	139	130
4	10	22	29	26	2	14	11	19	9	63	40	55
5	5	5	4	10	1	3	3	8	1	22	18	17
6	2	5	3	2	0	2	2	4	1	12	8	11
7	1	5	3	4	0	0	0	2	0	2	3	5
8	0	2	0	1	1	0	0	0	0	1	3	1
9	0	0	0	0	0	0	0	1	0	1	0	2
10	0	0	0	0	0	0	0	0	0	0	1	1
more	0	0	0	3	0	1	0	0	0	2	0	6
ave*	.347	.554	.474	.563	.172	.395	.357	.421	.312	.813	.720	.779

*ave is the sum of the number of fragment for each indexer divided by 3,072

Selected overall totals

Total of averages 36,864

Total fractional representation

0 24,861 67%

1 08,141 22%

2 to 10+ 03,862 11%

Average for EGAD 0.492 fragments per signature

Average for 15,000 (2.08 X EGAD) 1.020 fragments per signature

Frequency of non-zeros

with one fragment per signature 68%

TABLE 2

EGAD 7216 TRANSCRIPTS
12 DIFFERENT 2-BASE ANCHOR SEQUENCES
1,024 DIFFERENT 5-BASE INDEXING SEQUENCES
4 DIFFERENT ENZYMES: Sau3AI, Tsp5091, Tail, NlaIII

frag #	AC	AG	AT	CC	CG	CT	GC	GG	GT	TC	TG	TT
0	3020	2556	2741	2525	3487	2901	2970	2866	3072	2143	2262	2238
1	803	1030	936	1060	519	881	854	866	808	1117	1105	1076
2	200	343	277	342	75	233	200	257	167	491	437	462
3	52	112	96	103	11	54	49	61	32	199	180	187
4	11	30	33	36	2	17	15	25	14	85	65	80
5	6	9	6	18	1	7	6	11	1	33	25	22
6	2	7	4	3	0	2	2	5	1	19	12	13
7	2	5	3	4	0	0	0	3	1	3	4	6
8	0	3	0	1	1	0	0	1	0	2	0	2
9	0	1	0	0	0	0	0	1	0	2	0	2
10	0	0	0	0	0	0	0	0	0	0	2	1
more	0	0	0	4	0	1	0	0	0	2	0	7
ave*	.356	.568	.484	.584	.176	.400	.367	.436	.320	.830	.746	.793

***ave** is the sum of the number of fragment for each indexer divided by 3,072

Selected overall totals

Total of averages 49,152

Total fractional representation

0 32,781 67%

1 11,055 22%

2 to 10+ 05,316 11%

Average for EGAD 0.505 fragments per signature

Average for 15,000 (2.08 X EGAD) 1.05 fragments per signature

Frequency of non-zeros
with one fragment per signature 68%